www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

1. MOLECULAR STRUCTURE OF DNA AND GENETIC INFORMATION

1.1 Structure and stability of nucleic acids

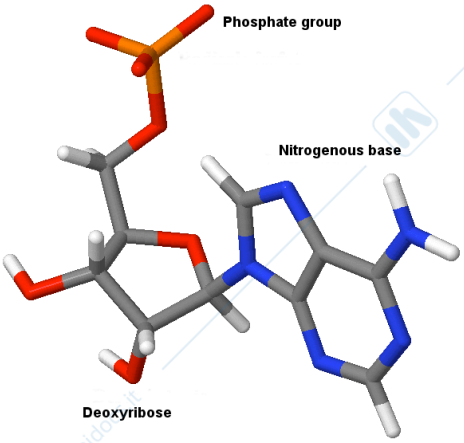


Figure 1.1. Adenosine monophosphate. Source: http://biomodel.uah.es/en/model3/nucleoti.htm

In 1953 James D. Watson and Francis H. Crick (Watson and Crick, 1953) proposed a model of the fine molecular structure of the ***double helix*** of deoxyribonucleic acid (***DNA***). Each of the two strands that make up the double helix is a nucleotide polymer. Each ***nucleotide*** (Figure 1.1) consists of a sugar molecule (deoxyribose, a pentose), a nitrogenous base (a basic compound rich in nitrogen) and a phosphate radical (ion of orthophosphoric acid).

DNA can be imagined as a polymer of this fundamental unit (polynucleotide), where there are two continuous chains that face each other. A single chain is composed of the constant alternation of a sugar and a phosphate group; from each sugar the nitrogenous bases, or more simply the "bases", are projected on the axis perpendicular to the direction of the chain. The two chains are perfectly analogous and face each other so as to place their own nitrogenous bases in the vicinity, towards the center of the space that remains between them, while the sugar-phosphate chains remain outside. Often the phosphate sugar chain is also called the "backbone" or skeleton because it can be thought of as a support from which the ends protrude (the bases).

A first feature to underline of the DNA is, therefore, that of possessing a simple, regular and periodic ***skeleton***, in which there are regularly, without exception, always and only two elements: a ***phosphate group*** and a ***sugar***. DNA differs from the other known nucleic acid, ***RNA*** (ribonucleic acid), mainly due to the sugar that is incorporated in it. In both cases it is a ***pentose***, that is a sugar with five carbon atoms, but while the DNA contains deoxyribose, a ribose lacking a hydroxyl (-OH) in position 2, RNA contains ribose, from which its name is also derived. The presence of hydroxyl in ribose gives RNA much less chemical stability since OH is a highly reactive group. This is also the reason why RNA degrades much more easily and quickly (it can be extracted intact only from viable cells), while DNA can remain almost unchanged for a long time, up to millions of years; this also allows its extraction from traces of degraded biological material, as in the case of investigative findings in forensic medicine and even from fossils.

Conversely, if RNA extraction from biological samples, such as from blood samples, is not done quickly with care, RNA will tend to easily degrade into low molecular weight fragments. RNA fragility and, conversely, DNA stability are also the reasons why genetic information is primarily conserved and passed on through DNA; despite this the DNA will be decoded in the primary structure of the proteins passing through a short-lived RNA intermediate. DNA can be compared to the volume of an encyclopedia, to a stable and permanent manual; RNA, on the other hand, can be imagined as a "work sheet", on which notes are transcribed from the main text, which is used for quick reference and then thrown away.

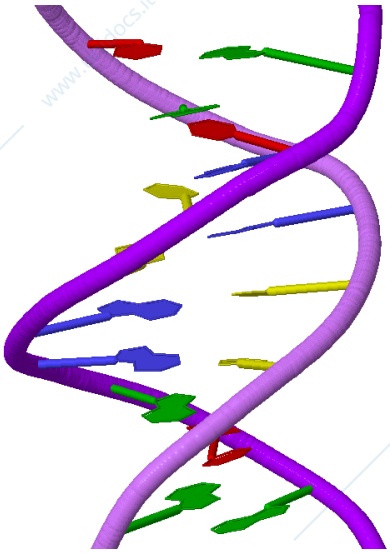


Figure 1.2. Simplified view of DNA structure. Violet: periodic structure. Blue: A. Green: G. Yellow: T. Red: C. Source: http://www.umass.edu/molvis/tutorials/dna/

1.2 DNA is an aperiodic polymer

The structure observed so far (in purple in Figure 1.2), the skeleton, is perfectly periodic, being characterized by the repetition of the sugar-phosphate-sugar-phosphate sequence that cannot carry information. To be an informational molecule, DNA must however be aperiodic (lacking in periodicity), because meaningful information is characterized by an irregular alternation of signals that say something precise.

A variable signal, according to the Watson and Crick model, appears instead on the arms protruding from the skeleton, where the ***nitrogenous base*** can be, in any position, one of the four different alternative bases (Figure 1.2): ***Adenine*** (***A***), ***Thymine*** (***T***), ***Cytosine*** (***C***) or ***Guanine*** (***G***). Therefore, the aperiodicity is given by this second structural characteristic of DNA; moreover the presence of a nitrogenous base rather than one of the other three in a certain position imposes no constraints from the point of view of the chemical structure on the choice of the base that follows it or precedes it along the same strand of DNA. The chemical synthesis of DNA has confirmed that a sequence at will is compatible with the double helix structure. According to the Watson and Crick model, there is instead a constraint: given the sequence of the bases on a strand, the sequence of bases in the opposite strand that wraps around the first will be determined; this is because, in the double helix, in front of each base on a strand, only a certain

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

base can find space and binding stability, according to fixed pairs of mutual recognition: Adenine pairs with Thymine, Cytosine pairs with Guanine (Figure 1.3).

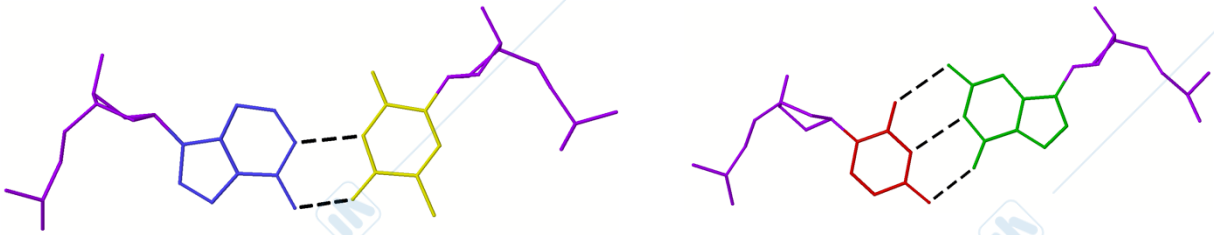


Figure 1.3. Matches between the bases. On the left pairing of an adenine (blue) with a thymine (yellow), to the right of a cytosine (red) with a guanine (green). Hydrogen bridges are highlighted in black. Source: http://www.umass.edu/molvis/tutorials/dna/dnacode.htm

When describing a DNA sequence, by convention we simply list the initials of the nitrogenous bases that follow one another along a strand, assuming with this that there is still a fixed skeleton of sugars and phosphates-and that the sequence of the undeclared strand is obtained for complementarity, that is according to the canonical pairings just described.

The structural basis for the fixed combination between the bases is first of all their steric bulk. In fact, adenine and guanine are larger bases, consisting of two rings of carbon and nitrogen atoms, while cytosine and thymine are smaller, being made up of a single ring. A and G are called Purines, C and T instead Pirimidines.

The ***historical names*** of these compounds are not related to their activity, but rather come from the name of the materials from which they were originally isolated: the names purine and pyrimidine derive from the Greek root "pir" (fire) because they were originally extracted from coal fossils and other combustible materials; cytosine was extracted from cells (cyto), thymine from the thymus (a gland), guanine from guano, strongly acidic, precisely because of its high content in nucleic acid derivatives, and adenine from glands (adenos in Greek).

The space between the two chains of the double helix is fixed and not sufficient to contain two large bases facing each other. Therefore, a pyrimidine is always found in front of a purine and vice versa. In addition, regardless of the strand on which they are found, adenine appears specifically with thymine; this happens because the two bases are able to establish between them two hydrogen bonds between the atoms that face each other more closely. Similarly, cytosine and guanine are able to establish three hydrogen bonds between them.
It follows that the separation of a pair of CG bases requires more energy than the separation of an AT pair.

The fundamental characteristics of the DNA structure are known from Chemistry and Biochemistry and are usefully summarized and tridimensionally visualized by a splendid tutorial available online:
http://www.umass.edu/molvis/tutorials/dna/

1.3 The concept of information

We are now dealing with how DNA is the substrate of genetic information. Information has the fundamental characteristic of communicating something that was not previously known; this lack of knowledge is generated by all the alternative answers, possible and equiprobable, for the original question. Information, therefore, from a physical point of view, appears as a ***choice***. Simply writing an A on an empty board represents information, because among all the possible signs that could be made, we chose to write A.
Information theory was developed by Claude E. Shannon in the 1940s, offering for the first time the tools to treat information as a physical quantity (Shannon, 1948).

Figure 1.4. Crossroads. Source:

The information can be defined as a ***reduction of uncertainty, which is obtained by sending or receiving a message***. If we are at a crossroads (Figure 1.4) and we do not know which way to go, receiving information on the right path reduces our initial uncertainty.
A first consequence of this definition is that ***if there is no uncertainty, there is no information***. If there had not been a choice between several possible alternatives there would have been no uncertainty, therefore, no need for information exchange. In uncertainty about a road to travel, asking for directions and receiving a response reduces the uncertainty of the route to take and, therefore, represents information. If there is only one road, there is no need for more information.

Figure 1.5. Roundabout. Source:

Secondly, the greater the uncertainty (Figure 1.5), that is ***the greater the number of alternatives, the greater the information*** provided through the choice of only one of the possible alternatives. In this perspective, it is clear that knowing, for example, the first amino acid of a protein (between the

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

possible winds) represents a greater amount of information than to know the first nucleotide (between four possibilities) of a nucleic acid sequence. Therefore, DNA contains information, such as proteins, but at a lower level of complexity. From this perspective we can also "quantify" the information contained in a single sequence. In fact, the identification of a known base on a DNA strand coincides with the choice of one of four alternatives, while the choice of an amino acid identifies one of twenty possibilities. We can measure the information in bits, which is the minimum unit of information and represents the amount of information associated with the choice between two possible alternative states (binary alternative: open-closed, on-off, 0-1). The name "bit" is derived, in fact, from the contraction of "**bi**nary digi**t**" in English. Therefore, in each known position on the DNA there are two bits of information, since the alternative between four possibilities is attributable to a double binary choice (alternatives 00, 01, 10 and 11). If these bases are associated in a chain, the information contained is greatly increased because, for example, a chain of 5 nucleotides has 1024 possible combinations (4 raised to 5). Furthermore, the transformation in figures of the information contained in the DNA allows it to be preserved and analyzed with specific tools.

1.4 DNA and information

In some ways, the DNA contains all the information necessary for the cells of an organism to position themselves and to function properly; the cells are able to read and interpret this information. We thus arrive at a central concept: in nature there has been, for four billion years, the first bacteria, something that man discovered only in the twentieth century, digital information (from "digit" which means numeral in English, in turn from the Latin "digitus" which means finger, because fingers are used to count integers). This type of information is contained and transmitted in the form of discrete signals, i.e. it can be traced back to numbers.

The example of an ***analog*** clock and a ***digital*** clock is clarifying. The analogue clock, by "analogy" with the time it represents, has hands that flow seamlessly, so that at any moment the position of a hand represents exactly the fraction of time measured. This representation of time is "analogous" to time, that is to the object of measurement. Instead, the digital clock uses numbers to represent the flow of time and, however accurate, it is a "jumping" procedure, at discrete levels, without possible intermediates. In this representation, time is reduced to numbers.

Not all sizes can be represented in a non-analogical way, that is, in a digital way; until the twentieth century no one had imagined that it was possible to digitally transmit genetic information, that is the entire constructive plan of man, like any other living being. In fact, instead, all this information is converted into numbers and not by man, but in nature, through a language that requires only four types of signals. Therefore, a molecule exists in nature, such that atoms inside are not limited to occupying a space describing the shape of the object (as in the case, for example, of the atoms of a stone), but at the same time transporting the interpretable signals, so much so that by reading them and interpreting them you can reconstruct a new organism (and a stone cannot do it). In the mid-seventeenth century, microscopists had intuited that information for the development of man had to be contained in the eggs or in spermatozoa, but at an analogical level. They had imagined the so-called ***homunculus*** (Figure 1.6), a microscopic but fully formed man, huddled inside the spermatozoon, waiting to grow. The idea was born from the observation that already an embryo of a few millimeters or a newborn is completely formed in the general body plan, similarly to a man of two meters; going backwards by analogy seemed logical. The complexity, therefore, was generated directly by that of the mother-organism; it was unthinkable to pass “low information density” that can be converted into bits through a relatively simple molecule such as DNA. It could not be imagined because we arrived at the concept of digital information only centuries later; furthermore, this type of information is associated with an arbitrarily established coded interpretation. For this reason, the explanation of how the emergence in nature of coding molecules, or molecules with sequences of signals that are a prelude to the formation of a new molecule, according to a defined code, remains a currently unsolved mystery of Biology.

Figure 1.6. Homunculus. Source: https://it.wikipedia.org

Given that-DNA is the substrate of biological information, it is important to try to define how its various structural aspects are associated with the treatment of this information. The development of the molecular biology of the gene has made us understand that there are precise relationships between certain structural features of double-stranded DNA and their relationships with information management. In other words, there are relationships between the form and function of DNA.

1.5 Structure and function of DNA in relation to the management of biological information

The ***structure-function*** relationship is fundamental in nature; at each level we observe that, if a thing has a certain form, it performs a certain function. A simple example: the stomach is made in a certain way, that is, it is hollow (while the liver is a full organ) because it is a container and must contain the food inside it. This observation, obvious at the macroscopic level, is valid in nature up to the microscopic and molecular levels; if a certain molecule

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

has some atoms protruding in a certain direction this will confer certain properties, for example of binding, to that molecule. Just as one structure recognizes another at the macroscopic level because they have complementary forms, so the whole functioning of the great biological molecules (macromolecules) is attributable to a principle which is the ***complementarity of form***: one form must be complementary to the other because these they can be hooked together. In addition to this condition, the interacting molecules must have ***bonds*** that stabilize this recognition. Note that the chemistry of interactions between biological molecules is characterized by the preponderant use of weak bonds (hydrogen bonds, hydrophobic interactions); these bonds offer a much greater plasticity and dynamism, more suited to the variability that characterizes living beings.

On the basis of these principles, all the recognitions between informational macromolecules take place: DNA-DNA, RNA-DNA, protein-DNA, protein-RNA, protein-protein, but also among some signals encoded in the sequence of glucose or lipid molecules.

The structure of DNA ultimately accounts for the way information is ***represented***, ***transmitted***, ***used*** (i.e. read and interpreted) and ***changed***. Let us now analyze these four aspects.

Genetic information is contained in DNA: the single strand

First of all the information must be present and the specific succession of bases in a single strand of DNA is sufficient to define it. In each position, we have one in four chances of accidentally having a certain base; every base we specify increases the informational content of the sequence. We can state that the aspect of "containing" information is already present in a single strand; the other strand has a necessarily complementary sequence. There is therefore no need, *a priori*, for a double strand. In fact, many viruses have a single-stranded genome, sufficient to carry information for the construction of viral proteins. The need for the double strand emerges from considerations linked to the repeatability and stability of the sequence (thanks to the "quality control" that each sequence exercises over the complementary one), rather than the recording of information as such. Furthermore, a strand can undergo changes independent of those of the opposite strand, as happens for example in the case of DNA methylation, which is asymmetrically distributed between the two strands (see the topic: Epigenetics), increasing the wealth of information of the molecule.

We also observe that larger amounts of information will be contained in longer molecules; in fact, in general, the simplest organisms possess genomes of smaller dimensions than more complex ones (Bainard and Gregory, 2013).

Genetic information is transmitted through DNA replication: the double strand

Several reasons could explain the presence in all non-viral genomes of a second strand of nucleic acid complementary to the first. One could think of a greater chemical stability, but usually the stabilization of DNA is due to the proteins linked to it, while the ability to form double helix stretches, more thermodynamically stable, is also a prerogative of single strands (intracatenary folding).

It is more reasonable to look for an explanation precisely in the informational aspects of the molecule: the presence of two complementary copies of the same sequence of elements represents reciprocal ***quality control*** of the sequence itself. If one of the two strands of the mold is damaged, the sequence can be reconstructed by reading information on the other strand, as they are, in effect, capable of doing DNA repair enzyme systems.
If there is information carried by a single-stranded sequence and there is a break in the glycosidic bond between the sugar and, for example, base A, I cannot trace the original information; if, however, there is a complementary strand with the base T at the level of the damaged position, during replication it will be possible to regenerate a new strand with the A in the correct position.

Probably, in nature, a principle evolved (4 billion years ago) that was discovered by engineers in 1700-1800, when they started to build machines that were more and more reliable from a functional point of view. For example, on a bicycle or car, there are usually two brakes because in case one breaks it is still possible to use the other. This simple principle is called ***functional redundancy*** ("two is better than one") and is valid in Biology at multiple levels, from molecular to anatomical; for example, many organs are vascularized from multiple vessels, so that it is still possible for blood to reach the organ in case of a vessel obstruction.

From a biological point of view, it is particularly important that the information present in a cell is correct when we transmit it to the daughter cells that can originate by division and, in particular, to the reproductive cells destined to form a new individual. Given the mechanisms of DNA replication, which work with the "complementary mold" principle, to obtain two identical sequences starting from an initial sequence it is necessary to have two strands. Each strand represents a biochemical "cast" of the other, therefore it is different on the one hand, but it is able to function as a mold, so as to obtain two indistinguishable double helices at the end of the replication.

Ultimately, the double-stranded structure is closely related to the problem of ***faithful transmission of information***. This is the second great function linked to the structure of DNA, which must not only contain information, but must also be able to transmit it; the genetic "text" is "photocopied" and checked before being passed on to another user.

Genetic information is selectively read and used: transcription and translation

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

The information contained in the project of a house is read and interpreted by the builders ~~up to~~ responsible for the physical construction of the building. Similarly, DNA is in effect digital information that contains a project of how an organism must structure itself and function, so the cell must contain the components necessary to use this information. However, this information must be used at the right time: just as building a house, it would not make sense to read the instructions to place a dividing wall when the foundations are not yet completed.

The structural aspect of the DNA, which ensures the possibility of selectively reading this information, is the recognizability of the atomic groups of each of the four bases; these groups are exposed to the outside (Figure 1.7) and are different from those facing the center of the double helix which establish hydrogen bonds with the appropriate atoms of the complementary base.

The fact that the sequence of bases can be "read" from the outside, like a specific configuration of the groups of atoms of the relative bases, without the need to carry out the double helix and to use the complementarity of the bases, allows some large proteins to "land" literally on the closed double helix, establishing links between specific atomic groups of protein amino acids and a determined sequence of bases on DNA. This event is associated with the subsequent opening of the double helix in the downstream section of the one bound by these proteins and to the RNA copying (transcription) of the sequence of bases of the performed region.

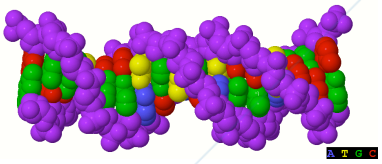


Figure 1.7. Structure of DNA. The atomic groups of the bases are exposed to the outside. Violet: periodic structure. Blue: A. Green: G.Yellow: T. Red: C. Source: http://www.umass.edu/molvis/

It can be said that these proteins ("transcription factors") select, at a given moment, a certain page of the project to copy and carry out the instructions written on that page.

Therefore, the third fundamental relationship between form and function tells us that the sequence of DNA bases can be recognized not only by base-base contact, but can also be identified from the outside by looking at the "back" of the bases from some proteins whose structure is able to specifically bind the atoms exposed externally by a particular sequence of bases. This coincides with the subsequent reading and use of the information contained in that region; in general, this information consists of the instructions for the assembly of the amino acids in a given polypeptide chain.

According to the "central dogma of molecular biology" the main function of DNA is to specify the production of the organism's proteins, through the intermediate of RNA (mRNA). Proteins make up most of the dry weight of a cell and are, therefore, fundamental components of the protoplasm; the body, however, also needs glucose and lipids, components of the same nucleic acids, as well as small molecules. Why, then, does genetic information focus on transmitting the characteristics that ***proteins*** must have and not those of other constituents?

The answer must explain both aspects of biological information, namely the transmission of a ***constructive plan***, so that a spider will have eight legs and a man two legs, and the transmission of instructions to manufacture the ***components*** of the organism. In the same way, the architect indicates the shape of the building of a project, i.e. the relative arrangement of the parts and the materials with which the parts will be made.

In what sense is it sufficient to specify the protein amino acid sequence to fabricate all the ***components*** of the body's cells? Because proteins can play two fundamental functions: on the one hand, they constitute the "***building blocks***" of cellular architecture, that is, they are necessary for the realization of cellular structures, as are sugars and fats; on the other hand, many proteins can act as ***enzymes***. In nature, enzymes, or biological catalysts that accelerate cellular chemical reactions, are normally proteins; it can therefore be stated that the chain of reactions for the assembly of a polysaccharide proceeds anyway thanks to proteins. In other words, a protein can make a sugar, but a sugar cannot make a protein. Therefore, transmitting information to manufacture proteins is equivalent to transmitting instructions to realize all the main types of molecular "bricks" at the base of the living substance: proteins, fats, sugars and nucleic acids; it should be noted that the enzymes also involved in nucleotide metabolism and those involved in the mechanisms of repair, replication and functioning of DNA and RNA are, chemically, proteins. Note that this is a central paradox in the management of genetic information, whose solution is linked to the problem of the origin of life: DNA manufacturing is required to make proteins, but DNA cannot function without specific proteins. One theory envisages an initial "RNA world" in which this intermediate molecule could simultaneously perform catalysis and information support functions; this hypothesis is based on the catalytic properties actually shown by the RNA in some circumstances.

We must now consider the ***constructive plan***, that is the morphological plan, the project, of where the macromolecules in the cell and every cell in the organism must go to locate.

Indeed, the amino acid sequences also perform the function of directing the position of the various components of each cell and directing cellular movements. The localization of the proteins themselves is often guided by a "signal" peptide that coincides with about twenty amino acids at the amino end of the polypeptide chain, while particular proteins secreted on the outside or exposed on the membrane influence the survival, proliferation, differentiation, migration and function of different cells in many ways. Already in the early stages of development the cells show specific movements and, from an apparently disordered mass, an organization gradually emerges that outlines the arrangement that the various organs will have. In recent years, a number of fundamental "switches" have been identified to trigger the development of some organs or components.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Genetic information can change over time: spontaneous mutations

Considering only the first three aspects, we would have an organism that contains information, transmits it faithfully and uses it; however, it would always give rise to identical individuals. Information, on the other hand, varies continuously giving rise to ***biological variability***.

The biological variability is distinguished in two great aspects, that ***of species*** and that ***of individual*** within the same species. Today we know the data on the percentage of base pairs that differentiates, for example, the man from the primate most similar to him, the chimpanzee (about 1%) or a human individual on the other (about 0.1 %). If the project varies beyond certain limits you can have a new species; if, instead, there are small variations there are different individuals of the same species. Once again all this is due to the DNA sequence. Small changes in the sequence mean that none of us is completely identical, genetically, to another individual (even in the case of monozygotic twins).

The molecular basis of the possibility of varying the sequence is mainly due to the chemical mode of recognition between the bases. For reasons related to the chemical-physical structure of the carbonaceous rings of the nitrogenous bases, it is possible that the redistribution of the charges, which occurs due to the oscillations of the electrons, involves the modification of the atomic groups involved in the pairing of the bases (the assumption of an alternative structure by a chemical compound is called tautomerism). These changes, if they occur during DNA replication, can allow wrong pairing. There is no perfect machine in nature; even the complex molecular machinery that copies DNA is therefore not error-free. Multiplying these errors over time gives rise, on a large scale, to variations to make new species or to change the mutant individuals of a certain species.

1.6 General organization of genetic information in humans

From the biological point of view, *Homo sapiens* is composed of apparatuses, the organs from organs, the organs are finally aggregates of tissues that are composed of cells. The fundamental unit of living organisms is the cell, a unit distinct from others and capable of living alone; it is bound by a membrane and has a nucleus that contains the vast majority of genetic information. A precise estimate of how many cells a human body has actually does not exist; estimates range from 10^{13} (10,000 billion) to 10^{14} (100,000 billion), which is probably more realistic. The most recent estimate of the number of cells of an "average" man is 3×10^{13} (30,000 billion), of which the great majority is made up of red blood cells, while the number of nucleated cells is estimated at 3,000 billion (Sender et al., 2016). Each of these cells has the same 46 chromosomes, 46 DNA molecules of the same sequence as those present in the zygote (fertilized oocyte), originating from the spermatozoon and the egg. Since then the individual is genetically homogeneous, with some exceptions that will be discussed in the context of the case of monozygotic twins.

The human genome (hereditary heritage) consists mainly of 46 chromosomes (DNA molecules of an organism): 46 double-stranded linear DNA molecules of which the largest is about 250 million base pairs (chromosome 1) located in the cell nucleus. Our genome is also divided into two sets of chromosomes similar to each other in pairs, one that derives from the father and one that derives from the mother for each pair. If we consider a single kit, the total length is about ***3 billion base pairs*** (3.2 Gbp, giga = 1 billion).

If we further enlarge the level of investigation, we see along the chromosome regions called genes, which alternate continuously with other ***regions*** that are not genes, called ***intergenic DNA***. From a molecular point of view the gene is defined by biochemists as a ***region of DNA that contains the information necessary for the synthesis of a polypeptide chain***. A polypeptide chain is a chain of defined sequence amino acids; the information to mount the amino acids in a protein in the correct order is contained in the DNA. In reality, only about 1/3 of human DNA is made up of genes coding for proteins (about 22,000 estimated), that is, it contains information that somehow makes it possible to have a protein, while the remaining 2/3 contain repeated sequences, much of viral origin.

Ultimately, in every cell there is information to fabricate all the proteins and to orient all cells within the constructive plan of the whole organism; however, in the individual, each cell uses this information only partially. The debate on this topic has lasted years; in the mid-1960s it was still unclear whether the differentiated cells gradually lost DNA segments. It was observed that a red blood cell was very different from a neuron and it was thought that they could not derive from the same genetic information.

Instead, surprisingly, all cells maintain the same information present in the zygote, but each "reads" only a part of the information, selecting it for the functional purposes characteristic of the cell itself.

The construction of an organism can be compared to that of a house, made up of many rooms (cells); the instructions for building this house are in an encyclopedia of 46 volumes (DNA), each bound volume is a chromosome. Each volume contains information to make a number of components. In this analogy, every room in the house contains an entire copy of the encyclopedia. The individual chapters of each volume explain how a certain component must be constructed and can be compared to genes; the single letters of the chapter, instead, are the single base pairs. The paper is a non-informative support, like the sugar-phosphate skeleton, and on it the letters are ordered that constitute the message in a precise sequence.

The beta-globin gene is "read" a lot by the blood cells, which draw the information to make many copies of the beta-globin, a constituent of one of the main red blood cell proteins, hemoglobin; the beta-globin gene is present in the same sequence even in brain neurons, however in this case it is a page that brain cells never read.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

When an individual reproduces, he transmits a halved copy of his genome, contained in the spermatozoon (for the man) or in the egg cell (for the woman); these kits merge into a new cell, the ***zygote***, and the DNA is initially read to get information on how to divide cells. However, by merging two haploid genomes of mammalian oocytes into the same cell, the embryo does not develop. This is because there are DNA methyl transferases, enzymes that transfer methyl groups to certain nucleotides, as a sort of chemical highlighting (epigenetic, what lies above genetics, analogous to the yellow highlights above the DNA sequence). The methyl transferases of a spermatozoon are different from those found in the oocyte. So, if a cell is masculine it will express a certain class of these enzymes and, consequently, it will attack the methyl groups at certain points. If a cell is feminine it will show different points. Thus, when there is sexual reproduction, nature requires that one passes through a male cell on one side which is a mirror image, complementary, like chemical marking, with respect to the female cell on the other side. Only the union between the two germ cells will generate an initial cell (zygote) with a functioning genome from which, at an incredible speed due to exponential doubling (for repeated somatic divisions, mitosis), the morphology of a subject is eventually formed as human; within the different cell types, specialized instructions are beginning to be used for all the individual components necessary for the function of that cell type. It is estimated that, on average, in a single fertilization in our species 200 million spermatozoa are brought into play. Of these, only one will bring its genetic program to work along with the oocyte genetic program. The real selection of how the individual is made comes from which sperm will first enter the oocyte, which in response will emit a calcium wave that will block the membrane for the entry of any other sperm.

When does the new individual start from a biological point of view? It was thought that since information is in the nucleus, the individual is there when he has only one nucleus. Then instead it was understood that the problem was the chromosomes and not the nucleus; therefore, the individual is there when he has 46 chromosomes. Therefore it is not necessary to wait for the two nuclei to merge, but even if the individual has 2 nuclei, that cell already has 46 chromosomes. Now there is the new bioinformatics paradigm for which ***the individual is when he has a complete, selected sequence that will work***. In a sense, it is when the first calcium atom comes out of the lattice to block the entry of other spermatozoa which determines the complete genetic information that will characterize the individual.

Among the individuals of one species and those of another, the variations are both at the level of structural proteins and at the level of the proteins that govern the constructive plan. However, functional "themes" common to organisms that are also far from each other emerge.

For example, does a fish have genes (and therefore proteins) similar to humans? Mostly yes, but with some important differences; the scale protein is a type of keratin, but different from the keratin of the skin or human hair: this is a difference in the cellular and tissue "bricks". There is also a second fundamental difference at the level of the construction plan: a fish has no arms and legs, but has fins; this information is already present in the zygote and is read gradually during the development of the individual, which from the zygote will lead to the formation of an adult individual recognizable as a fish. Even in this case, however, common elements emerge: the head is at the end, the process of developing the fins uses genes similar to those responsible for limb development, and so on.

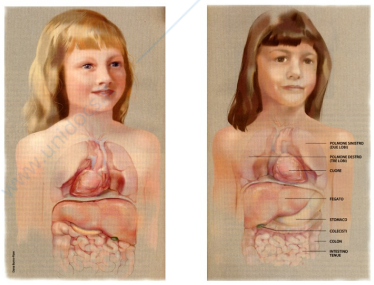


Figure 1.8. Situs viscerum inversus. Source: Le Scienze, August 1999.

That the constructive plan is in the DNA was definitively confirmed only in the 90s, when a genetic mutation which gives rise to a benign anatomical anomaly consisting in the inversion of position between the right and left side of the thoracic and abdominal viscera (*situs viscerum inversus*, Figure 1.8, OMIM #270100, http://omim.org/entry/270100) was characterized. Some genes encode proteins (dyneins) that function as small engines; these proteins are produced by cells on the center line of the embryo's body and make the eyelashes of cells always beat in the same direction, creating a "current" of molecules that establishes an asymmetry between left and right. If the DNA of one of these genes shows the wrong sequence, the protein does not work and the molecules will be arranged at random. This can provoke this rare condition where the organs are arranged asymmetrically, as in a body reflected in the mirror. Moreover, half the time, by pure chance, for example the left molecules will be placed on the right side and other times they will not be. In fact, half of the individuals who carry this mutation have the organs in the correct site, while the others have the mirrored site. This is the final and formal proof that the "project" of the human body form is written in DNA.

In a few decades it will be possible that you will have a complete list of all the genetic organ switches. For example, the genetic switch was discovered that first triggers the development of a tooth (*ENAM*, http://www.ncbi.nlm.nih.gov/gene/10117), so that individuals who have this mutated gene have complete edentulism.

The geneticist Jérôme Lejeune (1926-1994) argued: "*Each of us has a very precise starting moment which is the time at which the whole necessary and sufficient genetic information is gathered inside one cell, the fertilized egg, and this moment is the moment of fertilization. There is not the slightest doubt about that and we know that this information is written on a kind of ribbon which we call the DNA, It's a long molecule in which, under a specific code, all the qualities of the future persons are defined. It measures exactly 1 meter of length, split in 23 little bits*

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

inside the chromosomes, and 1 meter inside the ovule; so that, at the beginning of our life, we have got 2 meters of so to speak magnetic tape in which everything is coded". (Lejeune, 1990).

Summing up the gene is that region of DNA that contains information to make a protein; this protein is either a "brick" (plastic component) or a factory of other components (enzyme) or it is even the worker who indicates ***when*** and ***where*** the bricks must be produced (gene transcription factor) and how they must be mounted (regulatory proteins). Thus, it is sufficient to propagate the information that is in the DNA to "reconstruct" an entire individual; just as it is sufficient to deliver a project of a house already built to an engineer so that he can build a similar one at another point without starting from a miniature model. The digital information found in germ cells is sufficient to reconstruct an entire individual, but it must be reunited in a cell capable of interpreting it: the DNA isolated in a test tube cannot manufacture anything, just as a project that is not found in a shipyard with specialized workers remains inert. Life requires this unity of the digitized project (comparable to ***software***) and a living cell that has components that can use it (comparable to ***hardware***).

Bibliography

Bainard JD, Gregory TR. Genome size evolution: patterns, mechanisms, and methodological advances. *Genome* 56:vii-viii, 2013.
http://www.ncbi.nlm.nih.gov/pubmed/24168634

Lejeune J. Testimony before the louisiana legislature. June 7, 1990.
http://publications.fondationlejeune.org/article.asp?filename=fjl424.xml

Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. bioRxiv 2016.
http://biorxiv.org/content/early/2016/01/06/036103

Shannon CE. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27:379-423, 1948.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6773024

Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171:737-738, 1953.
http://www.ncbi.nlm.nih.gov/pubmed/13054692

Websites

Structure of DNA
http://www.umass.edu/molvis/tutorials/dna/

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

2. STRUCTURE AND FUNCTION OF THE EUKARYOTIC GENE

2.1 Gene concept

The subject of this discussion is not so much the biological and biochemical details of the functioning of the gene, but a definition of the concept of gene. It is so true that the functioning of the gene can be hypothesized on a conceptual basis, that the essential ideas of the working scheme of genetic material sometimes precede the relative experimental documentation, starting from Mendel. The term gene (from the Greek "genesis" = birth or "genos" = origin) was coined by Wilhelm Johannsen, Danish botanist, in 1909 (Johannsen, 1909): genes are "*special conditions, foundations and determinants that are present* [in gametes] *in a unique, separate and therefore independent way* [through which] *many characteristics of the organism are specified*" (Gerstein et al., 2007). Thus, in classical genetics, the ***gene*** was the ***basic unit that transfers genetic information in a manner consistent with Mendel's laws***.

Note, therefore, that the *gene* is a concept, not a separable physical object from the others. Indeed, even the demonstration of the physical substrate of the gene in the form of a "region" of a DNA molecule does not allow us to affirm the physical individuality of the gene, which appears to be linked to other sequences of the molecule on which it is located. Conversely, the chromosome appears naturally as a physical object, which can be, for example, "fished" and separated from the others.

Following the discovery that mutations in genes could cause defects in the metabolic pathways of the *Neurospora* organism (Beadle and Tatum, 1941), we have the well-known definition of "a gene, an enzyme", which later became "a gene, a polypeptide". The fact that inheritance had a molecular and physical basis was supported by the demonstration that the substance transmitted by bacteriophages to their offspring was actually DNA (whose double helix model was proposed by Watson and Crick in 1953) and not protein (Hershey, 1955).

The idea was that a ***gene***, a certain region of DNA, contained ***instructions for the synthesis of a protein***, which in turn controls a certain hereditary character.

Russian physicist George Gamow hypothesized that the amino acids adapt directly to the furrows of the DNA and, in this way, were aligned and welded together (Gamow, 1954). Instead, from the 1960s onwards, it was established that first an intermediate molecule called mRNA ("messenger RNA", or messenger RNA, because it carries the DNA message to the protein factory, or ribosomes) is produced. Why is this intermediate phase needed? The most reasonable explanation seems to be the possibility of selecting the genes to be activated allowed by this mechanism.

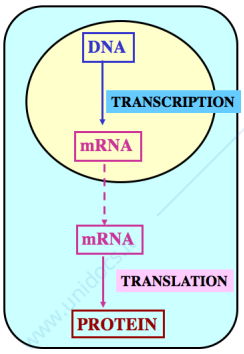


Figure 2.1. Scheme of transcription and translation processes in a cell. Redrawn by Lorenza Vitale.

Transcription is the process by which RNA molecules are synthesized, starting from the mold consisting of one of the two strands of DNA. The transcription process takes place in the nucleus (Figure 2.1). When RNA is synthesized, it can exit from nuclear pores and localize into the cytoplasm, where mRNA information is translated into proteins using the genetic code (solved by Nirenberg et al., 1965; Söll et al., 1965). This second process is called translation (Figure 2.1).

It is interesting to note that the terms chosen to indicate the fundamental processes of genetic information management are taken from linguistics, confirming the analogy between different types of code. ***Transcribing*** means copying a text from an original source to a new medium, in our case from the genome to a "sheet of notes" which is the mRNA. The genes active in a specific cell type and functional at some moments are those transcribed.

This synthetic text arrives in the "workshop", where it is used to direct the manufacturing of the piece (the protein). For this it is necessary to ***translate*** the instructions on the leaflet into a series of appropriate assembly operations, which lead to the final product. The gene product has the structure of an amino acid chain; therefore, the translation in the biochemical sense coincides with the synthesis, on the ribosomes, of a polypeptide chain, performed on the basis of the mRNA mold.

The ribosome is made up of RNA and proteins; the particular type of RNA found in the ribosome is called ribosomal RNA (rRNA); this RNA is the backbone of the ribosome and is involved in the catalyzing of the formation of the peptide bond. The other type of RNA, which is fundamental for translation, is called RNA transfer (transfer RNA, tRNA); it performs precisely the function of ***translator*** between genome language (4 letters, nucleotides) and the language of proteins (20 letters, amino acids). The fundamental characteristic of a translator is the knowledge of two different languages, combined with the ability to relate them to each other. Even before the discovery of tRNA, the existence of a generic "adapting molecule" was postulated, a molecular structure capable of simultaneously recognizing the two languages. An example of a common adapter can be that of an electric reducer, which is made like the tRNA: on the one hand it recognizes a certain type of electrical socket, for example the Italian one, and converts it appropriately into another type of socket able to hook up, for example, the German plug of the new refrigerator (Figure 2.2). The characteristics of the tRNA correspond to those expected from a molecular adapter: it possesses a portion (anticodon) capable of complementarily recognizing a specific

Figure 2.2. Adapter.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

sequence of three bases on the mRNA and an end able to bind a specific amino acid depending on the anticodon sequence. For example, CCC is the codon for proline; there will be a tRNA that with a stroke recognizes the three CCC bases (a short single-stranded GGG sequence), and at the opposite end it will bear a Proline, specifically hooked by a specific enzyme whose structure is able to simultaneously recognize the anticodon and the corresponding amino acid.

We should note that rRNAs and tRNAs are transcribed from specific DNA regions, but are not translated into proteins. Then, if a region of DNA can be the one that when activated it produces an rRNA, the classical definition of gene is already in crisis: Can we call the region of DNA that is not transcribed in an mRNA but in rRNA or tRNA a gene? Certainly these gene products are important for the phenotype, such as proteins; indeed, without them the synthesis of proteins could not take place. Therefore, a broader definition of the ***gene*** as a ***transcription unit*** has emerged in Molecular Biology. The gene is thus ***any DNA region that can be transcribed***.

In other words, if a particular DNA region is transcribed, then it is a gene. In some cases the product of the transcription is an mRNA, in other cases the product consists of rRNA, tRNA or other types of RNA that are not translated and do not fall into the previous categories (Figure 2.3). The latter are now grouped in a class of RNA defined as a class of non-coding RNAs (ncRNAs), whose function is still largely unknown and probably have a regulatory role on the transcription of other genes.

It is important to note that the DNA molecules of our genome are similar to pairs (***diploidy***), for which we possess two similar copies of each chromosome with the exception of the sexual ones in the male, one derived from the father and one from the mother. When we say "the gene of a certain protein", for example the gene for human beta-globin, we refer to that region of a chromosome that has the instructions to synthesize beta-globin. The beta-globin gene is located on chromosome 11, but since our genetic makeup is diploid, we have two copies of the beta-globin gene. From a qualitative point of view, it is sufficient that the information for the correct synthesis of the beta-globin chain is present only once. However, we are in the presence of a new case of functional redundancy: the presence of two copies of the gene allows that in cases of lesion of a copy, the other can still direct the formation of the required polypeptide chain. In fact, the vast majority of ***genes for proteins*** are present in ***two copies***.

The ***redundancy*** principle is applied even more in the case of rRNAs and tRNAs. These gene products are part of the fundamental protein synthesis machine for manufacturing all proteins. Thus, a severe mutation in a gene that is essential to synthesize the thousands of cellular proteins would be incompatible with life. For this reason, there are hundreds of copies of the genes for the various types of rRNA and tRNA, moreover scattered in different chromosomes. Human diseases due to specific lesions of these genes are not reported.

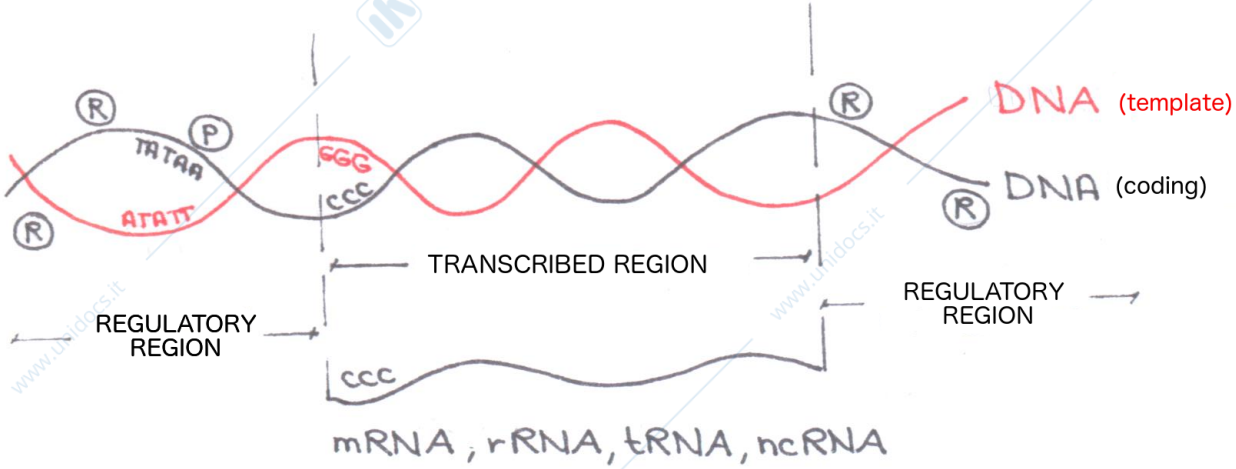


Figure 2.3. The gene. P: promotor; R: regulatory sequences. Illustration by Amalia Sanna Passino.

2.2 Basic functioning of a eukaryotic gene that codes for a protein

A typical ***eukaryotic gene*** includes:

Regulatory sequences: recognized by proteins capable of binding to DNA, recognizing the sequence "from the outside", with a closed double helix; include the promoter and "enhancer" sequences.

Transcribed sequence: the region that is actually copied in RNA; it is the region through which it can be said that a certain DNA sequence is a gene.

Promoter: region involved in starting the transcription process. It is usually a stretch of 100-200 bases located immediately upstream (i.e. in position 5′) with respect to the site of the beginning of the transcription. Very often it includes conserved sequences of the "TATA", "CAAT", "GC" type (when read on the coding strand) which are recognized by proteins activating transcription by RNA polymerase; they establish what will be the DNA strand to

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

be transcribed. The promoter is not transcribed. The integrity of the promoter is required for the assembly of the "general machinery" of the transcription in response to transcriptional activators.

Example - human beta globin gene (reported below). A "TATA" type sequence is found upstream of the transcription start site, from base 74 to base 77 (taaa, in red, lowercase). If we call the transcription start site +1, the position of the sequence box will be from -30 to -27 relative to that site. In position -75 there is a sequence of the "CAAT box" type (caat, from 29 to 32).

Enhancer ("Cis-acting enhancer of promoter function"): region that can greatly increase promoter efficiency by binding other proteins that favor transcription (Activators). Unlike the promoter sequences, the "enhancer" sequences can be found upstream or downstream of the transcribed region, even at a distance of thousands of base pairs from it, or inside it. The "enhancers" can act at a distance thanks to the folding of the DNA that brings these sequences into contact with the gene to be activated. Moreover, they can be indifferently in "sense" or "antisense" orientation.

Silencer sequences work the same way, but they bind proteins that are inhibitors of transcription instead of activators

Example - beta globin gene (Homo sapiens), *HBB*, http://www.ncbi.nlm.nih.gov/gene/3043
GenBank sequence n. NM_000518, http://www.ncbi.nlm.nih.gov/nuccore/NM_000518

Double-stranded DNA (the other is obtained by complementarity)

```
                                      -75
   1 ccctgtggag ccacacccta gggttggcca atctactccc aggagcaggg agggcaggag
                  -30                                   +1
  61 ccagggctgg gcataaaagt cagggcagag ccatctattg cttacatttg cttctgacac
 121 aactgtgttc actagcaacc tcaaacagac accatggtgc acctgactcc tgaggagaag
 181 tctgccgtta ctgccctgtg gggcaaggtg aacgtggatg aagttggtgg tgaggccctg
 241 ggcaggttgg tatcaaggtt acaagacagg tttaaggaga ccaatagaaa ctgggcatgt
 301 ggagacagag aagactcttg ggtttctgat aggcactgac tctctctgcc tattggtcta
 361 ttttcccacc cttaggctgc tggtggtcta cccttggacc cagaggttct ttgagtcctt
 421 tggggatctg tccactcctg atgctgttat gggcaaccct aaggtgaagg ctcatggcaa
 481 gaaagtgctc ggtgccttta gtgatggcct ggctcacctg gacaacctca agggcacctt
 541 tgccacactg agtgagctgc actgtgacaa gctgcacgtg gatcctgaga acttcagggt
 601 gagtctatgg gacccttgat gttttctttc ccccttcttt ctatggttaa gttcatgtca
 661 taggaagggg agaagtaaca gggtacagtt tagaatggga aacagacgaa tgattgcatc
 721 agtgtggaag tctcaggatc gttttagttt cttttatttg ctgttcataa caattgtttt
 781 cttttgttta attcttgctt tctttttttt tcttctccgc aatttttact attatactta
 841 atgccttaac attgtgtata acaaaaggaa atatctctga gatacattaa gtaacttaaa
 901 aaaaaacttt acacagtctg cctagtacat tactatttgg aatatatgtg tgcttatttg
 961 catattcata atctccctac tttatttttct tttatttta attgatacat aatcattata
1021 catatttatg ggttaaagtg taatgtttta atatgtgtac acatattgac caaatcaggg
1081 taattttgca tttgtaattt taaaaaatgc tttcttcttt taatatactt ttttgtttat
1141 cttatttcta atactttccc taatctcttt ctttcagggc aataatgata caatgtatca
1201 tgcctctttg caccattcta aagaataaca gtgataattt ctgggttaag gcaatagcaa
1261 tatttctgca tataaatatt tctgcatata aattgtaact gatgtaagag gtttcatatt
1321 gctaatagca gctacaatcc agctaccatt ctgcttttat tttatggttg ggataaggct
1381 ggattattct gagtccaagc taggcccttt tgctaatcat gttcatacct cttatcttcc
1441 tcccacagct cctgggcaac gtgctggtct gtgtgctggc ccatcacttt ggcaaagaat
1501 tcaccccacc agtgcaggct gcctatcaga aagtggtggc tggtgtggct aatgccctgg
1561 cccacaagta tcactaagct cgctttcttg ctgtccaatt tctattaaag gttcctttgt
1621 tccctaagtc caactactaa actgggggat attatgaagg gccttgagca tctggattct
1681 gcctaataaa aaacatttat tttcattgca atgatgtatt taaattattt ctgaatattt
1741 tactaaaaag ggaatgtggg aggtcagtgc atttaaaaca taaagaaatg atgagctgtt
1801 caaaccttgg gaaaatacac tatatcttaa actccatgaa agaaggtgag gctgcaacca
1861 gctaatgcac attggcaaca gcccctgatg cctatgcctt attcatccct cagaaaagga
1921 ttcttgtaga ggcttgattt gcaggttaaa gttttgctat gctgtatttt acattactta
1981 ttgttttagc tgtcctcatg aatgtctttt cactacccat ttgcttatcc tgcatctctc
2041 tcagccttga ct
```

***

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

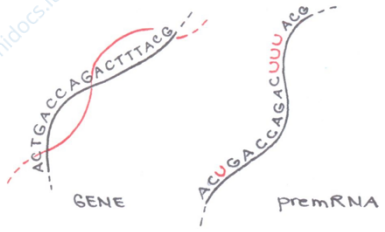


Figure 2.4. Differences between DNA primary transcript (pre-mRNA). The template DNA strand is shown in red. Illustration by Amalia Sanna Passino.

Primary transcript: the rough transcription product, consisting of an RNA strand that begins with the first transcribed base (site of transcription start, base +1) and ends beyond the polyadenylation signal.

The sequence of the primary transcript is complementary to that of the "mold" strand (*template*) of the DNA and corresponds to that of the strand opposite to the mold called "*coding*" or "sense", with the exception of the presence of the base U in place of the base T (Figure 2.4).

Example - human beta globin gene: the primary transcript is obtained by copying the mold strand with an RNA polymerase II from 104 to (at least) 1709.

Exon: ~~R~~region of the gene that codes for part of the mature mRNA (subject to splicing).

An exon is any gene sequence found in RNA at the end of maturation in mRNA, regardless of whether this sequence includes codons that will be translated into amino acids.

Example - human beta globin gene: EXON 1 from 104 to 245 (in color GREEN, UPPERCASE), EXON 2 from 376 to 598 (in color BLUE, UPPERCASE), EXON 3 from 1449 to 1709 (in color RED, UPPERCASE).

Intron (or ***intervening sequence - IVS***): ***transcribed region, excised during splicing***.

The sequence of the intron is, therefore, present in the primary transcript, interposed between two exons and is excised during the splicing process so that it is not found in the mRNA. The introns almost always begin with the sequence GT (GU in the pre-mRNA) and end with the sequence AG (classical or canonical introns). The sequence of a pre-mRNA is rich in AG and GU and the splicing complex understands which are the correct sites as they are surrounded by other signals.

Example - human beta globin gene: intron n. 1 from 246 to 375, intron n. 2 from 599 to 1448 (in lowercase, black; the intron start and end signals i.e. the splicing *donor site* and the splicing *acceptor site*, respectively, are in blue, lowercase).

pre-mRNA or Primary transcript of the beta-globin gene (Homo sapiens)
Single-stranded RNA

```
                                                  5´-ACAUUUG CUUCUGACAC
AACUGUGUUC ACUAGCAACC UCAAACAGAC ACCAUGGUGC ACCUGACUCC UGAGGAGAAG
UCUGCCGUUA CUGCCCUGUG GGGCAAGGUG AACGUGGAUG AAGUUGGUGG UGAGGCCCUG
GGCAGguugg uaucaagguu acaagacagg uuuaaggaga ccaauagaaa cugggcaugu
ggagacagag aagacucuug gguuucugau aggcacugac ucucucugcc uauuggucua
uuuucccacc cuuagGCUGC UGGUGGUCUA CCCUUGGACC CAGAGGUUCU UUGAGUCCUU
UGGGGAUCUG UCCACUCCUG AUGCUGUUAU GGGCAACCCU AAGGUGAAGG CUCAUGGCAA
GAAAGUGCUC GGUGCCUUUA GUGAUGGCCU GGCUCACCUG GACAACCUCA AGGGCACCUU
UGCCACACUG AGUGAGCUGC ACUGUGACAA GCUGCACGUG GAUCCUGAGA ACUUCAGGgu
gagucuaugg gacccuugau guuuucuuuc cccuucuuuu cuaugguuaa guucauguca
uaggaagggg agaaguaaca ggguacaguu uagaauggga aacagacgaa ugauugcauc
aguguggaag ucucaggauc guuuuaguuu cuuuuauuug cuguucauaa caauuguuuu
cuuuuguuua auucuugcuu ucuuuuuuuu ucuucuccgc aauuuuuacu auuauacuua
augccuuaac auuguguaua acaaaaggaa auaucucuga gauacauuaa guaacuuaaa
aaaaaacuuu acacagucug ccuaguacau uacuauuugg aauauaugug ugcuuauuug
cauauucaua aucucccuac uuuauuuucu uuuauuuuua auugauacau aaucauuaua
cauauuuaug gguuaaagug uaauguuuua auauguguac acauauugac caaaucaggg
uaauuuugca uuuguaauuu uaaaaaaugc uuucuucuuu uaauauacuu uuuuguuuau
cuuauuucua auacuuuccc uaaucucuuu cuuucagggc aauaaugaua caauguauca
ugccucuuug caccauucua aagaauaaca gugauaauuu cuggguuaag gcaauagcaa
uauuucugca uauaaauauu ucugcauaua aauuguaacu gauguaagag guuucauauu
gcuaauagca gcuacaaucc agcuaccauu cugcuuuuau uuuaugguug ggauaaggcu
ggauuauucu gaguccaagc uaggcccuuu ugcuaaucau guucauaccu cuuaucuucc
ucccacagCU CCUGGGCAAC GUGCUGGUCU GUGUGCUGGC CCAUCACUUU GGCAAAGAAU
UCACCCCACC AGUGCAGGCU GCCUAUCAGA AAGUGGUGGC UGGUGUGGCU AAUGCCCUGG
CCCACAAGUA UCACUAAGCU CGCUUUCUUG CUGUCCAAUU UCUAUUAAAG GUUCCUUUGU
UCCCUAAGUC CAACUACUAA ACUGGGGGAU AUUAUGAAGG GCCUUGAGCA UCUGGAUUCU
GCCUAAUAAA AAACAUUUAU UUUCAUUGC. ...-3´OH
```

* * *

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

mRNA: messenger RNA, derives from the primary transcript subject to a "maturation" process which includes the addition of the *cap* at the extreme 5′ to protect the end of the RNA from degradation, the addition of adenylic acid residues to the extreme 3′ and *splicing*.

At the end of maturation, the RNA (now mRNA) is composed of a ***series of exons welded together***. Three regions of mRNA are distinguished:

5′ untranslated region (5′UTR**, also called **Leader)

Stretch of the mRNA that extends from the *cap* up to the base that precedes the codon at the beginning of the translation (AUG).

Example - human beta globin mRNA: The 5′UTR region ranges from 1 to 50. Note that, in this case, the 5′UTR region consists of the first half of Exon 1; this does not exclude that it can also be composed of several exons in the case of genes in which the translation begins in exon 2 or 3 etc. *Capping* consists of adding a methylated guanosine in position 7 (m7G) upstream of the first transcribed base. This nucleotide is mounted so that the free end is a 3′-OH, so the mRNA has no free 5′ extremes.

Coding sequence (CDS), or "open reading frame" (ORF).

The CDS is the part of the mRNA whose sequence consists of a succession of triplets (codons) coding for the amino acid sequence of a polypeptide chain; it also includes the stop codon. The first three bases of the CDS are those of the codon of the beginning of the translation (usually AUG), while the last three bases of the CDS are those of one of the three codons of arrest (or stop) of the translation (UAA, UAG, UGA). Among these extremes is included a series of codons (on average 552 in the human genome) all coding for amino acids according to the rules of the genetic code.

Note on the definitions of CDS and ORF. In the biomedical literature there is no complete agreement on these definitions, which are often used without a formal declaration of their meaning. The prevailing idea, however, is that ORF is an "anatomical" concept ("open reading frame": row of codons between a start and a stop, "open", that is, without stop codons interrupting it), while CDS is plus a "functional" term: that potential sequence of codons is actually translated into a protein ("coding sequence").

In fact, for example, it is possible at the very least that very short sequences upstream of the main CDS called "Upstream Open Reading Frames" (uORFs) in turn present a start and stop codon, in this case according to some studies the corresponding small polypeptide could also be translated, although the main purpose of these small ORFs seems to be to adjust the translation speed. However, these are exceptions. A software could also look for ORF in a sequence, but without any experimental evidence that the ORF belongs to a transcribed and translated gene (in which case it would be an actual CDS).

Then, in practice, the two terms often appear to be used interchangeably. Also because, operationally, in the analysis of the sequenced genomes, the fact that a sequence can be a CDS is suggested by the "annotation" made on the sequence (at the mRNA level) that it is an ORF. The detection of ORF at the level of the DNA sequence is simple in the bacterial genes (without introns) and in the monoexonic eukaryotic ones, much more difficult in genes with more exons (and therefore subject to splicing) given the a priori unpredictability of the sites of splicing.

Example - human beta globin mRNA: the CDS ranges from 51 to 494; it is represented in UPPERCASE UNDERLINED. The translation start and stop codons are shown in bold (**AUG**, from 51 to 53; **UAA** from 492 to 494). The bases of CDS are 444, corresponding to 148 codons (444/3), of which 147 encode for the amino acids of beta-globin (including the initial Methionine) and 1 is the stop signal.
In this mRNA, therefore, the CDS consists of about the second half of Exon 1, all Exon 2 and about the first half of Exon 3. The number of exons in human genes can range from 1 to 363 (on average it is 11).

3′ untranslated region (3′UTR,** also called **Trailer)

Stretch of the mRNA extending from the first base that follows to the stop codon up to the poly-A tail.

Example - human beta globin mRNA: The 3′UTR ranges from 495 to the poly-A tail; it consists of the terminal part of Exon 3. The residues of A (about 50-250) added after the cutting of the primary transcript are represented in MAGENTA, which is carried out about 20 bases downstream of a specific ***polyadenylation signal*** (the canonical sequence is **AAUAAA**, in bold, at position 602-607 on the mRNA). The cutting point of the primary transcript, at the level of which the A residues are added, is called the ***polyadenylation site*** (in our case it is the base C at position 626 of the mRNA, hence the first non-A base before the poly-A tail).

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

beta globin mRNA (Homo sapiens)
Single-stranded RNA

```
    [3´OH-m7G]
  1 ACAUUUGCUU CUGACACAAC UGUGUUCACU AGCAACCUCA AACAGACACC AUGGUGCACC
 61 UGACUCCUGA GGAGAAGUCU GCGGUUACUG CCCUGUGGGG CAAGGUGAAC GUGGAUGAAG
121 UUGGUGGUGA GGCCCUGGGC AGGCUGCUGG UGGUCUACCC UUGGACCCAG AGGUUCUUUG
181 AGUCCUUUGG GGAUCUGUCC ACUCCUGAUG CAGUUAUGGG CAACCCUAAG GUGAAGGCUC
241 AUGGCAAGAA AGUGCUCGGU GCCUUUAGUG AUGGCCUGGC UCACCUGGAC AACCUCAAGG
301 GCACCUUUGC CACACUGAGU GAGCUGCACU GUGACAAGCU GCACGUGGAU CCUGAGAACU
361 UCAGGCUCCU GGGCAACGUG CUGGUCUGUG UGCUGGCCCA UCACUUUGGC AAAGAAUUCA
421 CCCCACCAGU GCAGGCUGCC UAUCAGAAAG UGGUGGCUGG UGUGGCUAAU GCCCUGGCCC
481 ACAAGUAUCA CUAAGCUCGC UUUCUUGCUG UCCAAUUUCU AUUAAAGGUU CCUUUGUUCC
541 CUAAGUCCAA CUACUAAACU GGGGGAUAUU AUGAAGGGCC UUGAGCAUCU GGAUUCUGCC
601 UAAUAAAAAA CAUUUAUUUU CAUUGCAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
    AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA ...3´OH
```

***

Amino acid sequence of beta-globin (Homo sapiens)
Polypeptide – one-letter code

Sequence obtained by translating the coding sequence (CDS) of beta-globin mRNA according to the rules of the genetic code:

(HBB, 147 amino acids, http://www.ncbi.nlm.nih.gov/protein/NP_000509):

```
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

***

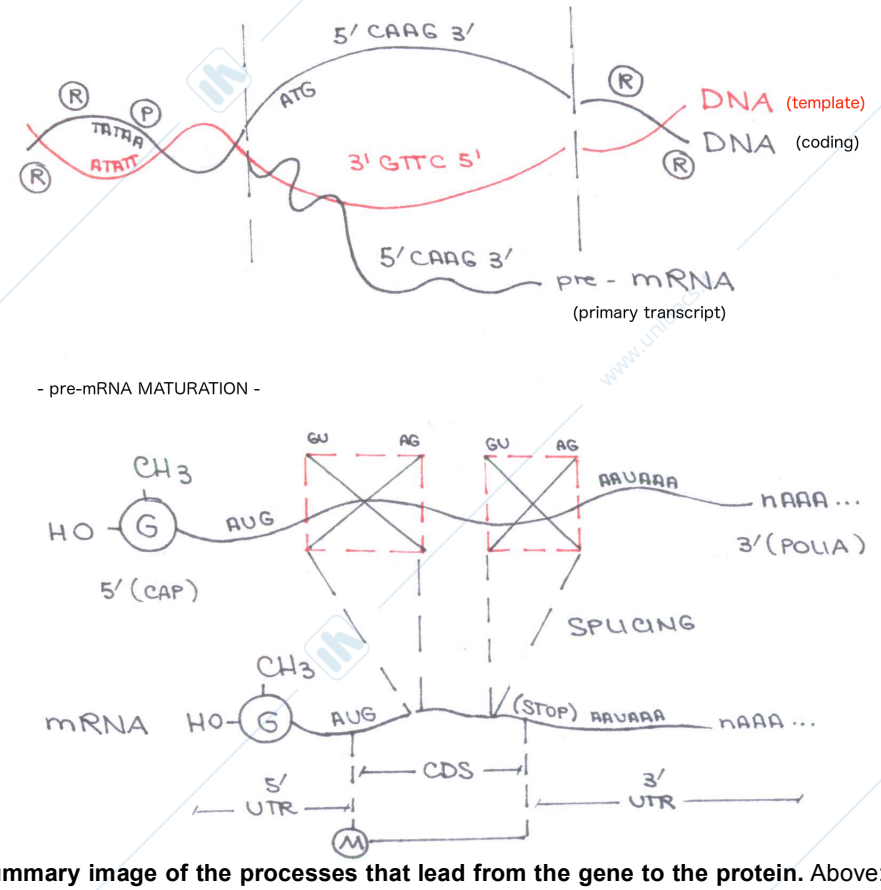


Figure 2.5. Summary image of the processes that lead from the gene to the protein. Above: schematization of the structure of a gene and of the production of the primary transcript (pre-mRNA). The promoter (P) and the regulatory sequences (R) remain closed double helix and are not transcribed. Bottom: the maturation of the pre-mRNA to obtain a mature mRNA by adding the *cap* to the 5′ terminal end, adding residues of adenylic acid to the 3′ end and splicing; the coding sequence (CDS) is then translated into an amino acid sequence starting from the translation start codon (AUG) which codes for methionine (M), up to the termination codon (stop). Illustration by Amalia Sanna Passino.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

2.3 Recent evolution of the gene concept

The detailed study of thousands of genes, in particular those of humans, conducted in recent years has provided a huge amount of data on the plasticity of gene transcription and translation phenomena, leading to the disruption of the classical scheme: a gene, an mRNA, a chain polypeptide. Indeed, starting from the same locus (the position of the gene on the chromosome), more alternative gene products can be generated. We mention below the main phenomena described, bearing in mind that a given gene may not use any of the following mechanisms or may use any combination of them:

Figure 2.6 Alternative transcription. P: promotor. Illustration by Amalia Sanna Passino.

Alternative transcription - alternative use of different promoters. Depending on the region that is recognized as a promoter, the transcription can start from different points in the same region, thus generating different transcripts in their extremity 5′ (Figure 2.6).

This phenomenon affects at least ***52% of human genes***; in 17% of cases in a tissue-specific manner. This process can cause changes in the amino acid sequence at the amino end of the protein product of the gene.

Alternative splicing - alternative use of different splicing sites. Depending on the sites used as splicing points, primary transcripts that have the same sequence could be processed differently from one another (Figure 2.7). In this way, a set of mature mRNAs of different sequence is generated from the same gene. Thus, we define exon, a sequence that remains in the mature RNA transcript; consequently, the definition of intron and exon cannot be definitive but is only functional, operative.

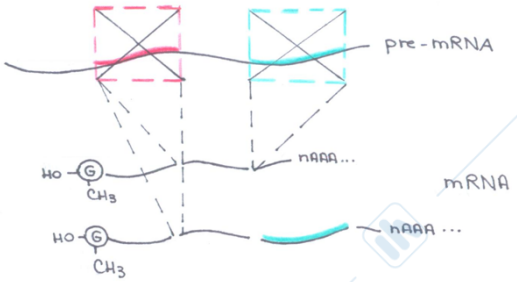


Figure 2.7. Alternative splicing. Illustration by Amalia Sanna Passino.

If the sequence difference between the mature mRNAs involves the coding region, the different types of mRNA will give rise, after translation, to different proteins. The different proteins originated by alternative splicing often have similar, but not identical, sometimes even opposite functions.

40-60% of human genes undergo alternative splicing and, on average, each human gene produces 3.5 different mRNAs. Homo sapiens are the species with the highest alternative splicing rate.

An effective example of alternative splicing, of importance in medicine, is the case of a hematopoietic growth factor (Figure 2.8): GCSF factor (*Granulocyte Colony Stimulating Factor* - factor that stimulates the formation of granulocyte colonies; official gene symbol: *CSF3*, *colony stimulating factor 3*, http://www.ncbi.nlm.nih.gov/gene/1440) which stimulates the formation of neutrophil granulocytes. Normally in human blood there are 50 to 70% of neutrophil granulocytes; calculating approximately 6,000 white blood cells per 1 µL, the presence of approximately 3,000 neutrophils is estimated. If there are 300, that is 10 times less, as unfortunately happens to some children who are born with a congenital defect of granulocytopoiesis and serious problems occur. To produce the right number of granulocytes, the GCSF factor stimulates the GCSF receptor (*Granulocyte Colony Stimulating Factor Receptor*, official gene symbol *CSF3R*, *colony stimulating factor 3 receptor*, http://www.ncbi.nlm.nih.gov/gene/1441), which is exposed on the membrane of granulocyte precursors and stimulates them to proliferate and mature in a granulocytic sense. It has been discovered that, in many cases, growth factor receptors are also produced in an alternative variant without the trans-membrane tract, a hydrophobic trait that is inserted into the phospholipid double layer and anchors the receptor to the cell surface.

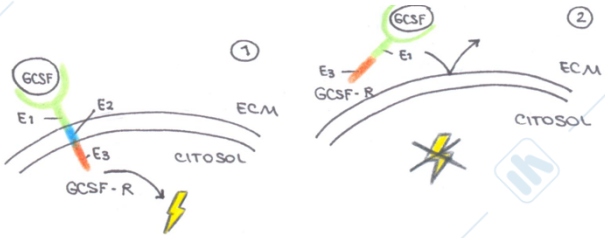


Figure 2.8. Example of alternative splicing. GCSF: granulocyte colony stimulating factor; E: exon; ECM: extracellular matrix; GCSF-R: receptor for GCSF. Illustration by Amalia Sanna Passino.

Imagining that for some splicing in some transcripts the exon encoding for the trans-membrane tract amino acids is removed (Figure 2.8, case 2), the protein will be produced without the triplets that encode this trait and therefore will not remain bound to the membrane. The produced receptor is called a soluble receptor as it comes out of the cell and is solubilized in the surrounding environment. When a factor binds to a soluble receptor it will not be able to trigger a cellular intracytoplasmic response since it is no longer bound to the cell; furthermore, the soluble receptor seizes the factor and prevents it from binding to the functioning receptor inserted in the cell. Thus the functioning GCSF receptor is on the membrane and triggers a proliferative and maturation signal, whereas

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

the soluble receptor not only does not trigger a signal but seize the signal; obviously a cell that produces an excess of soluble receptors cannot grow adequately. The gene can therefore produce two proteins with two opposite functions depending on the alternative splicing it adopts.

Alternative splicing could be compared to film editing; in fact splicing, in cinematographic English, means "editing" (literally "cut-and-sew"): the primary transcript is formed by all the scenes that the director has shot; the physical film, analog, is then cut into frames and stitched up to get the right scenes. Just as in the film the scenes can be mounted in a certain sequence, so in the messenger the exons can be mounted differently to obtain different proteins. In the same way, just as the insertion or not of a certain scene can change the meaning of the film, the insertion or not of an exon in the messenger can change the function of the encoded protein.

Figure 2.9. Alternative polyadenylation. Illustration by Amalia Sanna Passino.

Alternative polyadenylation - alternative use of different polyadenylation sites, consequent to the recognition of one of several AAUAAA polyadenylation signals present in the 3´UTR region of mRNA (Figure 2.9). Depending on the sites used as points of addition of the poly-A tail, primary transcripts that have the same sequence could have different sequences in the 3´UTR region.

About ***54%*** of human genes give rise to different mRNAs by this mechanism with possible consequences on the functioning and stability of the different mRNA isoforms thus generated. In fact, the 3´UTR region influences not only the stability of the transcript, but also the localization of the mRNA; for example, some proteins, which recognize this region, bind to microtubules and use them as a binary moving the mRNA to a certain point in the cell where the translation will take place.

Alternative translation - alternative use of different translation initiation sites, following the recognition of one of several AUG codons present in the 5´UTR region of mRNA.

Usually, in eukaryotes, the first AUG is always used starting from the 5´ end of the mRNA ("first-AUG rule"); this is because, unlike what happens in prokaryotes, the ribosome does not assemble directly "on" the beginning codon, but the minor subunit binds to the "cap" at the extreme 5´ and flows along the mRNA (***scanning mechanism***) until stopping at the level of the first AUG encountered, where the other components are assembled for the beginning of the translation.

The initiation codon in eukaryotic mRNAs is often found within a characteristic sequence, described by Marilyn Kozak (***Kozak sequence***: GCCGCCRCCAUGG, where the underlined bases represent those of the beginning codon) (Kozak, 1999). The most conserved features of this sequence, in eukaryotic genes, are the presence of a G in position +4 (considering +1 the base A of the beginning codon AUG) and / or the presence of a purine (R, i.e.: A or G) in position -3 (Figure 2.10). Usually, the closer the sequence surrounding the start codon is to that of Kozak, the greater the efficiency of mRNA translation.

In some cases, AUG codons subsequent to the first may be in an optimal sequence "context" for the start of the translation, so in a fraction of the mRNAs such codons will be used by the ribosome instead of the first one encountered; depending on the sites used for the start of the translation, mRNAs that have the same sequence could, in this way, give rise to polypeptide chains of different sequences at the level of the amino end. The phenomenon affects about 15% of human mRNAs.

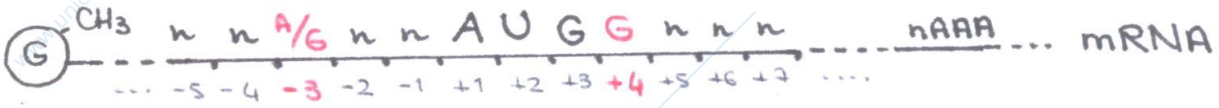


Figure 2.10. Kozak sequence. Illustration by Amalia Sanna Passino.

Localization and alternative function. The functional outcome of the different sequences of the gene products (mRNA and polypeptide chains), obtained by the previous mechanisms, can result in a different localization of the protein and / or in a different function; this is due to the fact that, in different cellular compartments, the protein can bind to different proteins giving rise to complexes that perform different functions. Most human proteins perform their function by binding to others; some estimates derived from ***proteomic*** research (study of large-scale proteins) indicate that, on average, a protein has binding sites for four other protein***s***.

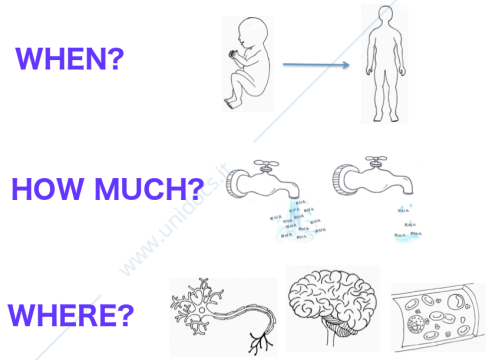


Figure 2.11. Expression of genes. Illustration by Valentina Serpieri.

The phenomena described have some important consequences for the ***redefinition of the molecular concept of a gene in the genome***.

First, it is noted that the physical boundaries of the gene are indeterminable from an "anatomical" point of view (in the sense of a specific "DNA segment"), and are identified by the methods of activation and processing of the gene which can vary at different functional moments or between different cell types.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Furthermore, the ability of a gene to supply different products, and according to variable schemes, has led Gelbart to affirm: "*I believe we are entering a period in which we must shift to the view that the genome largely encodes a series of **functional RNAs and polypeptides** that are expressed in characteristic spatial, temporal, and quantitative patterns. The classical concept of gene ultimately forms a barrier to trying to understand the phenotypes in terms of encoded functional products*" (Gelbart, 1998).

In practice, a gene alone can do nothing; it can express its information only in the context of the entire genome, in turn contained in a cell. If we wanted to describe exactly how a single gene works, we should take into account all the proteins needed to replicate and transcribe it. Following the traces of these proteins, we would notice that they perform their task having been produced following the activation of the genes that codify them and so on up to retrace the whole genome. There is an internal regulation in the cell, a real network, for which some genes are activated whose product is a protein that activates others, in a cascade process. In attempting an analogy, the genetic "text" appears in effect to be structured as a ***hypertext***, that is "***a structured set of information, united by references and links***". As in the *World Wide Web* of the Internet, the most complex hypertext realized, starting from a page, can virtually go through the whole network, in the genome, following the functional relationships of a gene to the end, one would find all the genes listed. 10-15% of our genes encode transcription factors, i.e. the molecular switches that activate other genes in an extremely coordinated way. Genes that must be activated together, because their products perform similar functions required in a given moment of cellular life can have similar regulatory sequences; in this way, the presence of a factor that activates one of these genes will also result in the activation of the other members of the "battery".

Since the type and number of gene products varies continuously in a cell, it is necessary to know the function of a normal or mutated protein not only in general, but within a specific cell type, i.e. in the specific operating context represented by all other proteins present in that cell type. A gene can be opened or closed, turned on or off (these are all equivalent terms). Gene on / on: it is transcribed; gene off / off: not expressed, not transcribed. Limiting ourselves to genes coding for proteins, within the estimated total of about 22,000 loci in the human genome, only about 19,000 are categorized as "known genes" namely genes well characterized and recorded in databases. It is possible to describe the status of any human cell indicating, for each known gene, whether it is transcribed or not. With modern methods able to study the global gene expression profile starting from cellular RNA it is in fact possible to determine, in less than 24 hours, which genes of a tissue were turned on and which were turned off at the time of RNA extraction; this generates an incalculable and inconceivable variability of possible cellular states: if each gene can have two states (on and off) and the genes are 19,000, the number of possible states is $2^{19,000}$ (the atoms of the whole universe are 10^{85} so if not already it is possible to think of the number 10^{85}; $2^{19.000}$ is not even conceivable for the human mind). Obviously the situation is complicated by adding the "very open gene" or "very closed gene" variable (Figure 2.12). For example, genes for hemoglobin are very "bright" in the precursors of red blood cells, so that hemoglobin forms 30% of the mass of a mature red blood cell; in the same red blood cell, before it loses its nucleus, the gene for insulin is also present but it is off. The **dynamic range** of the level of gene expression, i.e. the interval between the minimum and maximum levels of observable expression, can be of the order of magnitude of 1 - 100,000 or even 1 - 1,000,000, so a gene can show a variation of its very large, finely modulated expression intensity up to having circumstances in which the expression is a million times greater than the observed minimum level, as evidenced by the methods of studying gene expression by high-yield RNA sequencing (RNA-Seq).

The genes turned on / off vary from tissue to tissue: one neuron will turn on some genes, transcribe some genes, while a neutrophil will transcribe different ones; we speak of ***specific tissue expression***. Furthermore, many genes are transcribed (active) only during development and only in certain body regions. For example, there are genes expressed only between the third and fourth month of fetal life, only in a certain brain substructure, which then will never be expressed and are therefore turned off.

Ultimately, we should know precisely how many and which proteins are present in a certain cell (which in turn is in a certain functional stage and at a given moment in its development) and how they interact. The paradigm that is emerging today in the scientific community therefore expands the traditional "central dogma" of molecular biology (DNA → mRNA → protein), proposing a global view of the cell and its events, hence from the ***genome*** (complex of DNA molecules of a cell) a ***transcriptome*** (the set of all the transcripts of a cell) is dynamically expressed, from which following translation the ***proteome*** is obtained (the set of all the proteins present in the cell).

--

What is a gene? (from E. P. Fischer, 1995; modified)

If we asked...	**Would answer...**
A classical geneticist	The unit of transfer of genetic information according to Mendel's Laws

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

A biochemist	A region of DNA that contains information necessary to produce a protein
A genetic engineer	A fragment of DNA that can be introduced into one cell generating a phenotypic effect
A molecular biologist	A region on the chromosome that can be transcribed
A member of the Human Gene Nomenclature Committee	A DNA segment that contributes to the phenotype/ which has a function
Un geneticist of the post-genomic era	A region on the chromosome that can be transcribed generating families of gene products

GENE: TRANSCRIPTION UNIT (REGULATORY REGION + TRANSCRIBED REGION)

GENE: TRANSCRIBED
NOT TRANSCRIBED
=
ACTIVE
SILENT
=
EXPRESSED
NOT EXPRESSD
=
"ON"
"OFF"
=
SWITCHED ON
TURNED OFF
***

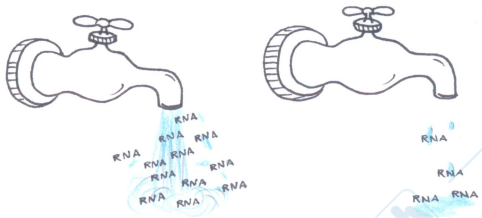


Figure 2.12. Active genes and silent genes. Illustration by Amalia Sanna

The concept of specific tissue expression also implies the fact that, if a given tissue is not analyzed at a given time, some transcripts may not have been studied yet and, therefore, it is not yet known that the corresponding DNA sequence is a gene. Therefore it cannot be excluded *a priori* that a given DNA sequence is not a gene and the number of genes cannot be exactly determined.

2.4 Evolution of the concept of "genetic code"

A code is an arbitrary set of rules that connect two independent worlds. An example of a code is language, which allows us to move from the world of objects to the world of words.

The g***enetic code*** makes it possible to predict, starting from a certain sequence of bases located on the mRNA, what the product of the translation will be (i.e. what will the amino acid sequence of the polypeptide chain be). However, this allows us to understand the operation of the only actual codifying bases, that is those that will constitute the codons on the mRNA: a percentage lower than 3%. We have not deciphered, for example, the t***ranscription code***, that is a relationship defined between the sequence of a promoter and the sequence of proteins that can recognize it, activating the gene; nor do we know the ***splicing code***, which would allow us to predict in what context a sequence can be interpreted as an exon rather than an intron. The ***folding code*** should allow us to predict the three-dimensional wrapping of a polypeptide chain, knowing its primary sequence.

All these codes actually exist in the cell, but we ignore their logic which is indispensable for the realization of genetic information. It is more correct to say that the code we have called "genetic" is actually nothing more than "one" of the codes used for the expression of genetic information and in this case is the ***translation code***.

The theory of existence in the biological world of multiple "organic codes" was developed by Prof. Marcello Barbieri, in particular in his book "I codici organici".

2.5 Methodological evolution in genetics

The methodological steps followed to arrive at the physical characterization of a gene have undergone drastic changes over the decades, for mainly technological reasons.

Traditional genetics

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Until the 1970s, methods for studying proteins in the laboratory were well developed; therefore, it usually started from a purified protein fraction which possessed a certain "activity", and an amino acid sequence, often partial, of the isolated polypeptide chain was obtained. At this point small "probes" were designed of sequence DNA compatible with that of the identified amino acids, and **hybridization** (pairing between two strands of nucleic acid having complementary sequence) made it possible to trace the corresponding gene region on the genome and sequence it. The operational flow that was used can be summarized as follows:

PROTEIN
SEQUENCING OF AMINO ACIDS
ISOLATION OF THE GENE
DNA SEQUENCING

Reverse genetics

Following the simplification of DNA sequencing, thanks to the introduction of the Sanger method (1977), it soon became easier to start from DNA and ***cDNA*** ("complementary DNA", that is a ***DNA molecule that represents a copy of an mRNA***) sequencing and to deduce the amino acid sequence according to the genetic code and the rules of translation ("first-AUG" and Kozak sequence). It should be noted that the amino acid sequences known as products of human genes were not, in large part, directly determined, but deduced from the mRNA sequence, thanks to the stability of the genetic code. This experimental strategy is called "***reverse genetics***" and can be summarized as follows:

ISOLATION OF THE GENE
SEQUENCING OF DNA AND cDNA
DEDUCTED AMINO ACID SEQUENCE
PROTEIN

Bioinformatics

In the mid-1990s the new method of "gene hunting" was established based on the availability of DNA sequences obtained on a large scale in the "genome projects", thanks to which the effort of systematic sequencing was taken on by a small group of specialized laboratories, coordinated at an international level. The public availability of the complete sequence of chromosomes and that of many mRNAs also systematically characterized in the "EST" projects ("expressed sequence tag", fragments of expressed sequences), allows the study to begin directly from a sequence of bases without having to determine it in your own laboratory. Through bioinformatic methods (computer analysis of sequences: comparison between DNA and mRNA sequences, comparison between different species, presence of a series of coding codons not interrupted by stop signals, etc.) the possibility is defined that a certain sequence code for an mRNA, which, if actually existing, will be later cloned in the laboratory easily (by retrotranscription and Polymerase Chain Reaction or PCR), based on the presumed sequence from the virtual reconstruction. From the sequence of the real mRNA we will then deduce that of the protein, which must be functionally characterized by the methods of ***post-genomics*** or ***functional genomics*** (inhibition of protein function in cells or organisms to observe the effect, protein research that binds to the one to be characterized, functional assays based on the presumed function based on the sequence similarity to proteins of known function, etc.). The stages of this strategy are therefore:

SEQUENCE OF DNA AND cDNA (AVAILABLE)
AMINO ACID SEQUENCE (DEDUCED)
PROTEIN STUDY

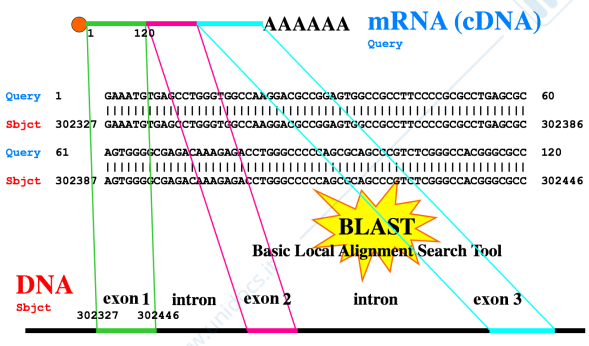


Figure 2.13. Representation of an analysis through the BLAST program. Explanation in the text. Illustration by Lorenza Vitale.

A bioinformatic sequence analysis is carried out using the BLAST program (*Basic Local Alignment Search Tool*, http://blast.ncbi.nlm.nih.gov/Blast.cgi), which in English means bursting, so called because it supplanted all previous sequence comparison algorithms. In short, when I use BLAST, I use a query sequence and the "launch" against a database that collects all the sequences of a certain type (for example all the nucleotide sequences known so far). The schematization of a result of a research carried out with BLAST is represented in Figure 2.13. The first line is the query sequence (*Query*), in this case a sequence of an mRNA, the one below is the sequence found (*Subject*). The perfect identity occurs from base 1 to

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

base 120 (Figure 2.13, green fragment of the mRNA) of the interrogation sequence and we see the correspondence with the foundations of the sequence found. Other results not shown in the figure will be useful to visualize the correspondence of the other fragments of the mRNA (pink and blue). By recording the positions with respect to the DNA sequence on the sequence found, I see that the fragment of green color is equal to the first fragment of the mRNA, that the second fragment of the mRNA corresponds to the pink fragment and that the third corresponds to the blue fragment. This allows us to understand that there are exons in the DNA, among which there are introns.

An in-depth analysis of the possibilities of bioinformatics is available on the website of our Laboratory: http://apollo11.isto.unibo.it/Flight_Manual/Bioinfo.htm.

References

Beadle GW, Tatum EL. Genetic Control of Biochemical Reactions in Neurospora. *Proc Natl Acad Sci U S A* 27:499-506, 1941.
http://www.ncbi.nlm.nih.gov/pubmed/16588492

Gelbart WM. Databases in genomic research. *Science* 282:659-661, 1998.
http://www.ncbi.nlm.nih.gov/pubmed/9784119

Gamow G. Possible relation between deoxyribonucleic acid and protein structures. *Nature* 173:318, 1954.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. What is a gene, post-ENCODE? History and updated definition. *Genome Res* 17:669-681, 2007.
http://www.ncbi.nlm.nih.gov/pubmed/17567988

Hershey AD. An upper limit to the protein content of the germinal substance of bacteriophage T2. *Virology* 1:108-127, 1955.
http://www.ncbi.nlm.nih.gov/pubmed/13267980

Johannsen, W. Elemente der exakten Erblichkeitslehre, Jena, 1909 citato da Nils Roll-Hansen. The crucial experiment of Wilhelm Johannsen. *Biol Philos* 4:303-329, 1989.
http://link.springer.com/article/10.1007%2FBF02426630

Kozak M. Initiation of translation in prokaryotes and eukaryotes. *Gene* 234:187–208, 1999.
http://www.ncbi.nlm.nih.gov/pubmed/10395892

Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, Rottman F, O'Neal C. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc Natl Acad Sci U S A* 53:1161-1168, 1965.
http://www.ncbi.nlm.nih.gov/pubmed/5330357

Söll D, Ohtsuka E, Jones DS, Lohrmann R, Hayatsu H, Nishimura S, Khorana HG. Studies on polynucleotides, XLIX. Stimulation of the binding of aminoacyl-sRNA's to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids. *Proc Natl Acad Sci U S A* 54:1378-1385, 1965.
http://www.ncbi.nlm.nih.gov/pubmed/5325653

Watson JD, Crick FH. Genetical implications of the structure of deoxyribonucleic acid. *Nature* 171:964-967, 1953.
http://www.ncbi.nlm.nih.gov/pubmed/13063483

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

3. THE HUMAN GENOME

3.1 Definition

The ***genome*** is the ***complete set of genetic information of a cell***. Note that when we talk about the genome ~~then~~ that we talk about the genetic information contained in a single real cell, we cannot talk about the genome in the abstract outside the reality of a cell that contains its own genetic information. In common language and in the media it is not rare to find some errors in the definition of the genome.

The most frequent error is perhaps that of equating the genome with the "genetic code", for example in the statements "it is written in our genetic code", or again: "the mutation is an error of the genetic code", because the genetic code is a set of rules that defines the translation of the coding triplets (codons) into amino acids ("translation code"), and as such cannot be wrong. It is as if the rules of grammar were confused with the text: the genetic code is by analogy the set of grammatical rules used to interpret a text, the genome is a specific text whose meaning can be understood according to the rules codified in the grammar of the language in which it is written.

It is also common to consider the genome as a "set of genes": in reality genetic information also includes areas of the genome that apparently are not genes, that is to say they are not transcribed or in any case no function is defined. However, it is always information represented in the form of a sequence of bases in DNA. From this point of view, independently of our knowledge of the functions of a given DNA sequence, it is correct to define the genome "biochemically" as the ***total mass of cellular DNA***. In fact, if the genome is the complex of the genetic information of a cell, but the genetic information is written in the form of a sequence of DNA bases, we can then say that the genome coincides with all the cellular DNA. Therefore, all the DNA molecules contained within a cell, not only those located inside the nucleus, constitute the genome: since there are DNA molecules that are outside the nucleus, such as the mitochondrial DNA, they must be considered fully part of the genome. Hence the correction of another error that actually makes the "genome" coincide with the "nuclear genome" alone, neglecting the genome component located in the cytoplasm.

One could object that there are viruses that have the ***RNA genome*** when discussing the universality of the DNA structure of the genome; in this case, which represents an exception, the definition "biochemistry" should be generalized stating that the genome is the total mass of the fundamental ***nucleic acid*** of that organism, even if a virus cannot properly be considered a living organism.

Finally it can be observed that since an individual can be constituted by one or more cells, when an individual transmits his characteristics, he does so by transmitting the information written in his DNA. From this point of view we can also consider the ***genome*** as the ***hereditary heritage of the organism to which it belongs***.

Structural genomics aims to provide a complete physical description of the genome, the genes it contains and the related gene products. ***Functional genomics*** aims to establish the functional role of individual genes, their interactions and their role within the organism to which they belong. In a broad sense we can say that the purpose of functional genomics of biology itself is to provide a complete physical description of the genome, the genes it contains and the related gene products. Functional genomics aims to establish the functional role of individual genes, their interactions and their role within the organism to which they belong. In a broad sense we can say that the purpose of functional genomics is that of biology itself.

The systematic treatment of the human genome can be articulated in the following scan: a first group of topics concerns the ***complexity of the genome***, of which we will analyze, with regard to the human species, the ***general organization*** and the ***variability***, as well as the ***number of genes*** and their ***classification***; then we will discuss ***gene families*** and ***pseudogenes***; finally the repeated sequences which, although present in the human genome, ***are not genes***.

3.2 Complexity of the genome

If the genome contains instructions for fabricating the components of the organism and directing their assembly at various levels, in this "engineering" sense it can be considered the equivalent of a manual describing the use of a machine. The manual of a machine that is not very complex is actually smaller than the one that describes a very complex machine, so we might expect that with the increase in the complexity of the organisms their genome will always be larger (Figure 3.1).

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

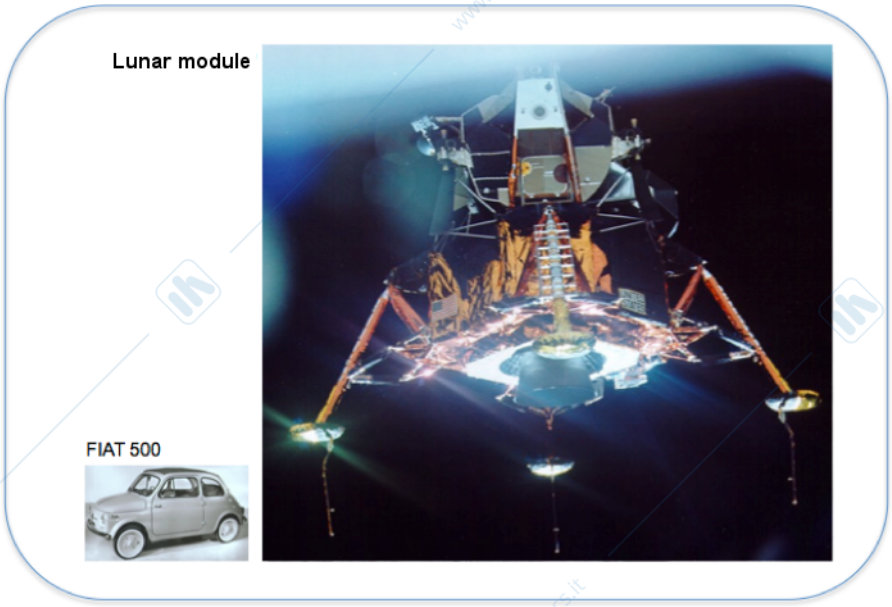


Figure 3.1. Simple and complex machines. The manual for the use of a subcompact like the FIAT 500 has 192 pages, the one for the LEM (Lunar Module) has 804 pages only for the subsystems. (The links to the files with the original documentation can be found on the site: http://apollo11.isto.unibo.it/appunti.htm)

Starting from ***viruses*** (Figure 3.2): the range of dimensions of the known viral genomes ranges from 859 base pairs (bp) (*Circovirus SFBeef*, single-stranded DNA genome or ssDNA) to 2.473.870 bp (*Pandoravirus salinus*, double-stranded DNA or dsDNA); for example, the single-stranded RNA genome of the HIV-1 virus is 9,181 nucleotides long (GenBank record No. NC_001802) (Campillo-Balderas et al., 2105). As an order of magnitude, we therefore go from 10^3 to 10^6 bp. Although according to updated data the largest viral genomes can overlap with the size of the smallest bacterial genomes, the ***bacteria*** genome has a range of orders of magnitude from 10^5 (*Nasuia deltocephalinicola*, 112.091 bp) to 10^7 (*Sorangium cellulosum*, 14.782.125 bp) (Land et al., 2015). ***Fungi*** genomes can be extended from 10^7 (*Hansenula polymorpha*, 8,97 Mb) to 10^8 (*Cenococcum geophilum*, 177,57 Mb) (Mohanta e Bae, 2015).

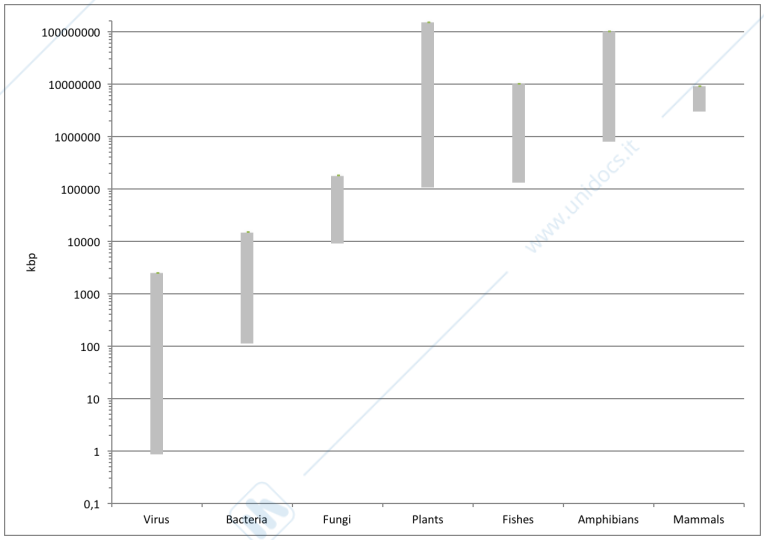


Figure 3.2. Paradox of the C value. Minimum and maximum size (in kbp) of the genome in viruses and in different groups of organisms.

While from viruses to fungi a clear increase in the average size of the genome is actually observed (Figure 3.2), from plants onwards this generic correlation between the complexity of the organism and the dimensions of its genome appears completely lost: ***mammals***, in particular man in whom we would expect the largest dimensions ever, have genomes of the order of magnitude of some billions of bp (10^9) but some ***amphibians*** can exceed one hundred billion bp (10^{11}) as well as some ***plants***. For example, the haploid onion genome covers about 16 Gbp (1 Gbp = 1 billion base pairs), resulting five times larger than the human genome. Very large genomes are described even among fish, mollusks and some insects. The absolute primacy is up to the *Paris japonica* plant whose genome reaches 150 Gbp; larger sizes described in the past for some amoebas seem to have not found confirmation with more precise modern methods.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

The lack of correlation between genome size and complexity of the organism was called in the past ***Paradox of the C value***, since *C-value* historically indicates the "constant" value ("c" stands for constant) or "characteristic" of the content haploid DNA found in the cell nucleus of an organism of a given species (Eddy, 2012). In the 1950s it was precisely the observation of the constancy of the DNA content in different cell types of the same organism, measured in pg (where 1 pg roughly equals 1 Gbp), contributing to the concept that DNA was responsible for inheritance. Notwithstanding the difficulty of the quantitative definition of the complexity of an organism that can then be correlated to a measurable parameter that in turn quantitatively describes the genome, it was decided to overcome this paradox by hypothesizing that the complexity correlates with other variables not always directly proportional to the genome size, for example the number of known genes present. However, today it is not possible to find any single variable that grows continuously with the increasing complexity of the organism; for example, the very number of genes can be very large in apparently not very complex organisms: in the rice genome (*Oryza sativa*) there are about 36,000 protein-coding genes compared to the approximately 20,000 human genes of the same class.

According to Eddy (2012) the paradox can be explained by observing that only a fraction of eukaryotic genomes is conserved and is functionally important and this is what increases with the complexity of the organism, while the rest of the genome is not functional. In this regard it was also stated that the size of the genome in eukaryotes is correlated to the quantity of repeated elements (transposons, see below) that constitute the genome itself (Elliott and Gregory, 2015). This explanation disagrees with the interpretation of the data of the ENCODE project (http://genome.ucsc.edu/ENCODE/) which attributes a functional role to 80% of the human genome. In this sense the complexity could be explained by the different functional mechanisms typical of the genome of a species (for example the phenomenon of alternative splicing is present at the highest level in human cells and in particular in neurons), therefore by the quality of information rather than quantity. The problem remains open.

3.3 General organization of the human genome

We will now discuss the general organization of the human genome (Strachan and Read, 2004). On the human genome sequencing website run by the public consortium is a screen that schematically shows the possible types of human chromosomes sorted by their size (from 1 to 22), the two sex chromosomes X and Y and the mitochondrial genome (Figure 3.3). Clicking on the corresponding graphic elements you can enlarge the chromosome until you read the basic sequence by base and have all the details for each chromosome.

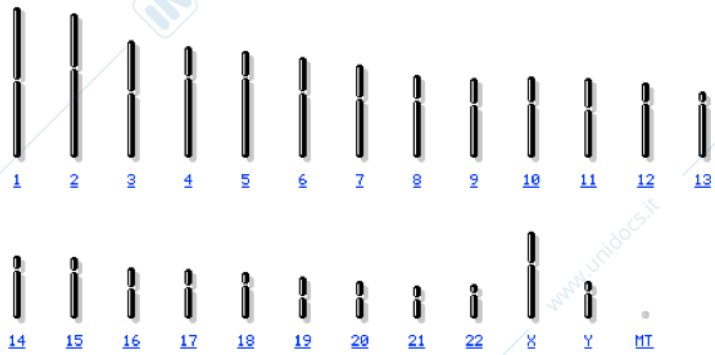


Figure 3.3. General organization of the human genome. Each type of human DNA molecule is thus represented in the "Map Viewer" web site (http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9606) of the *National Center for Biotechnology Information* (NCBI).

Some quantitative parameters of the human genome

To have a concrete and non-abstract idea of the human genome it is necessary to memorize some ***numbers*** that serve as a reference: the overall size of the human genome is usually indicated by referring to the human ***haploid*** genome, i.e. the DNA contained in a germ cell, so this amount will be doubled in a diploid somatic cell. Overall, the human genome is about ***3.2 billion base pairs (3.2 Gbp)***; therefore a diploid hepatocyte will contain 6.4 Gbp, but if it were a binucleate hepatocyte we would expect 12.8 Gbp. The human mitochondrial genome (***mtDNA***) is circular and ***16,569*** base pairs (***bp***) long, so it's enormously smaller than the nuclear genome. However, while the cell nucleus is usually single, mitochondria are present in many copies, on average from 800 to 2500 in a mammalian cell. Knowing that in the mitochondria there can be from 0 to 11 molecules of mitochondrial DNA (Cavelier et al., 2000), the mitochondrial genome is on the whole about ***0.5%*** of the total DNA, and its mutations can cause human pathologies, in particular neuromuscular.

Within linear nuclear DNA molecules, the longest is c***hromosome 1***: it contains about ***250*** million base pairs (***Mbp***). The last estimate amounts to 248.956.422 base pairs. ***Chromosome 21*** contains about ***50 Mbp***, so it is 5 times smaller than the larger chromosome. It is the smallest ever: according to the numbering of chromosomes,

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

which should reflect the progressively decreasing size of each type of molecule, it should have been chromosome 22, but due to an initial error in estimating the length of human chromosomes it was preferred to keep the classic numbering. It must therefore be emphasized that, between one human chromosome and another, there may be a difference in size of even ***5 times***.

To get an idea, chromosome 1 is about 8 cm long, although the double helix of DNA is only 2 nm thick. Adding the length of all types of chromosomes (only one ***haploid*** set) we therefore obtain the length of ***1 meter*** of DNA, therefore equivalent to about 3.2 Gbp. Since 1 Gbp of DNA weighs on average about 1.1 pg (1 picogram = 10^{-12} g = 1 billionth of a milligram), the total DNA weight is ***3.5 pg*** in a haploid cell (7 pg in a diploid cell). We are talking of a molecular structure of 2 nm in diameter!

In human diploid cells, the total length of 2 meters therefore includes a meter of paternal origin and a meter of maternal origin, and from the first cell of the organism (zygote) there will be 2 meters of DNA which are the exact sum of the maternal and paternal molecules contained in the respective gametes.

Recent estimates suggest that there are on average around ***30,000 billion cells*** in the ***human body***, of which about 3,000 billion are nucleated cells (Bianconi et al., 2013; Sender et al., 2016); assuming that they are largely diploid, we can estimate the total DNA length of a human organism to be around 6,000 billion meters, or 6 billion km, equivalent to about 40 times the Earth-Sun distance of 150 million km.

Distribution of different types of sequences in the human genome

About **32**% of human genome sequences are ***genes for proteins*** (transcribed and translated) (Figure 3.4), but since the codons actually translated into amino acids make up only a part of the gene, altogether the portion translated into protein of the entire genome human is about 1%: the other sequences are introns, which are removed from the primary transcript and are therefore not present in the mRNA and regions present in mRNA but not translated at 5′ (*5′ untranslated regions*, 5′UTR) and at 3′ (3′ *untranslated regions*, 3′UTR).

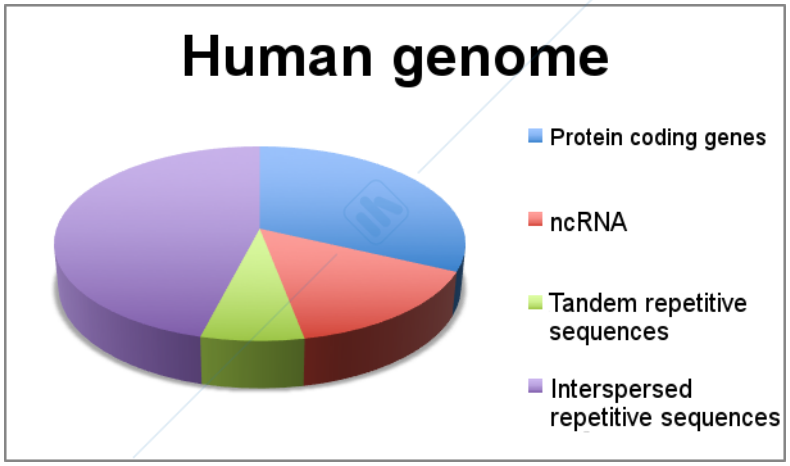


Figure 3.4. Percentage of different types of DNA sequences in the human genome.

The human genome is thus ***15%*** occupied by ***genes for non-coding RNAs*** (RNAs that are not translated, noncoding RNAs or ncRNAs), i.e. tRNA, rRNA and other regulatory RNAs. There are also pseudogenes in the human genome (genes that have stopped working) and there are non-genic sequences that are present only once in the genome. So far we have covered about half of the genome and wonder what is in the rest.

For ***7%*** these are ***repetitive sequences,*** one after the other ***in tandem***: for example, CAGCAGCAG...

46% is occupied by repetitive sequences not one behind the other, but present in different points of the genome where the sequence is always repeated. These are called ***interspersed repetitive sequences*** and it has been discovered that they are ***transposons***, sequences that are able to move their position from one point to another in the genome. As we shall see, it is believed that in man they have almost completely lost this capacity, otherwise they would continually create mutations. There are four classes in all: ***LINE***, ***SINE***, ***LTR*** and ***DNA transposons***, and may, though rarely, play a role in human genetic disease.

Map of the human genome

In 2001, the completion of the determination of the nucleotide sequence of human chromosomes allowed to put together the various types of maps that had accumulated over the years. It is a historical reference point for Genetics to have succeeded in ***aligning the microscopic structure of the chromosome***, i.e. a colored stick visible under a microscope with a series of light and dark bands after staining, ***with the sequence of DNA bases***.

For example, chromosome 11 is about 135 Mbp long (Figure 3.5). It contains more than 1,300 protein-coding genes, some of which are represented in Figure 3.5. We see for example that the gene for insulin (*INS*) is located towards the distal end of the short arm, in the sub-band 11p15.5 (the primary transcript extends from the nucleotide 2.159.779 to that in position 2.161.341, about 2, 2 Mbp) while the gene for beta globin (*HBB*) is found more

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

proximally, in the sub-band 11p15.4 at the height of 5.2 Mbp. The progesterone receptor gene (*PGR*) is found in the long arm, sub-band 11q22.1, in the coordinates around 101.1 Mbp. We can therefore accurately correlate the ***DNA sequence*** with the coordinates that mark the beginning and end of the genes present (determined by comparison with the RNA produced by the gene as the position respectively of the first and last transcribed base) and with their ***position on the chromosome***.

Note that a given gene can be transcribed from one strand or the other of the chromosome: in this sense we speak of ***gene orientation***. Of the two strands that the DNA possesses, some genes use a certain filament, and for example in Figure 3.5 the direction of their transcription will proceed upwards, others will instead use the opposite filament and then the RNA polymerase will perform the transcription proceeding towards the bottom.

Each gene has its specific direction in which it is transcribed (either upwards or downwards, i.e. from the 5´-3´ filament or from the antiparallel strand). The origin of the orientation of the various genes that follow one another along the chromosome in one direction or the other without a regularity in the type of orientation is not known; it can only be noted that overall about 50% of the genes per protein in the human genome is transcribed by a strand and the other half by the opposite one. By convention, each chromosome is represented with the short arm pointing upwards while the sequence that is actually recorded in the database is that of the strand in the 5´-3´ direction from top to bottom (plus strand or "+"). According to this convention, the orientation of the transcription for each gene is defined "+" if it occurs from top to bottom, indicating that the minus or "-" strand (the complementary one, not registered because it can be deduced from the other) is acting from a mold, and vice versa. When the gene is in "-" orientation, in the database its sequence carries the "complement" annotation. Of the genes mentioned above as an example, we observe that *INS*, *HBB* and *PGR* are all in "-" orientation, while the thrombin gene (coagulation factor II, symbol F2) is in "+" orientation.

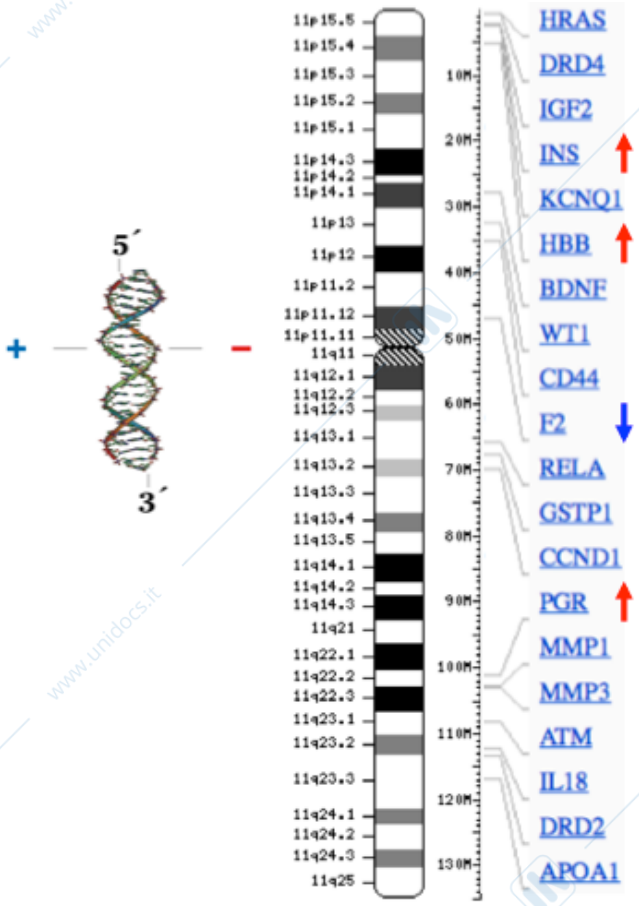


Figure 3.5. Human chromosome 11 map.
Source: *NCBI Map Viewer* web site.

The general concept of a ***map***, as a ***graphic representation of a territory***, suggests an analogy between genetic maps and geographic maps. The photograph of the chromosome under the microscope is equivalent to the photograph of a geographical region taken from the satellite where the physical structure of the object is shown as follows: 'is (physical card). The precise mapping of the boundaries of genes is in a certain sense equivalent to a political map, which shows the boundaries of the various states. With the completion of the human DNA sequence it was possible to overlap the two maps, just as in a physical-political map the territorial boundaries are superimposed on the natural elements.
In summary, just as in geographic maps we have four cardinal points, to orient a DNA sequence we must take into account two perpendicular "axes" that respectively define the ***position*** of the sequence given by the nucleotide coordinates along the chromosome (from the "highest" to the "lowest") and its ***orientation***, determined by which of the two "sides" of the double helix acts as an encoding filament, for example.

3.4 Variability of the human genome

The human genome is the genome of a particular cell and therefore of a certain human individual. One wonders to what extent the sequences of two human individuals are the same or different. The similarity between two human genomes must be sufficient to justify that organisms with a human genome are all *Homo sapiens sapiens* and all have the same general anatomo-physiological normal structure (two lungs, same type of blood circulation, same metabolism of the same substance ...), while on the other hand we expect DNA sequence differences that

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

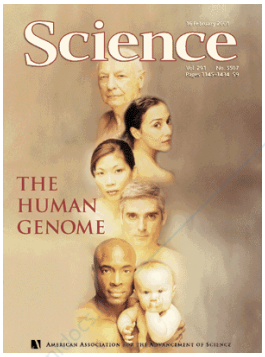
account for individual variability for all the characters that distinguish us and that have at least in part genetic basis (conformation of face, height, color of eye and hair, skin, predisposition or not to certain diseases...). In other words, what *percentage of difference between two complete human genomic sequences can simultaneously justify our similarity as members of the same species and our irreducible individual diversity*? An initial answer to this question was possible only after the sequencing of a complete human genome made possible by the "*Human Genome Project*" (*HGP*).

Human Genome Project

The sequencing of the human genome certainly represents a fundamental turning point in the history of Genetics (Ciccodicola and D'Urso, 1998), although it should be noted that it is not so much a question of advancing biological thought as of a technological showdown. In fact the premises for such a result, both conceptual (the genome is the complete sequence of DNA bases), and technique (it is possible to localize the position of a sequence on the chromosome), had already been formulated in the 1960s, an era in which science fiction writer and science writer Isaac Asimov predicted that the complete map of the human genome would be obtained (Asimov, 1968).

A concrete project to carry out the complete sequencing of human chromosomes took shape in ***1990***. The work was expected to be completed in 15 years; in fact, the first draft was released after 11 years, in ***2001***. In ***February 2001***, two articles were published describing the complete sequence of the human genome. The sequence published in the English journal ***Nature*** (International Human Genome Sequencing Consortium, 2001) in the historical issue dedicated to the theme "*The Human Genome*", was determined by a consortium of universities and public bodies (20 centers in 6 different countries) led by Francis Collins whose policy was the immediate release of data gradually obtained in online databases to be freely accessible to anyone. Important updates of this map were made available in 2004 (International Human Genome Sequencing Consortium, 2004), 2009 and 2013, the year that dates back to the latest version available, which has the abbreviation "hg38".

The sequence published by the American journal ***Science*** (Venter et al., 2001) was instead obtained by a private company, Celera Genomics, led by Craig Venter, which followed a methodological approach different from that of the public consortium and initially made data available only by payment.



It is now natural to ask which specific individual the published sequences belong to: this is a question with profound implications of ***genetic privacy***, because by making the subject studied known, there would have been strong pressure to allow its DNA sequence to be correlated to all its normal and pathological features. To "mask" the sequence, the public consortium used the DNA extracted from eight blood samples obtained from 5 men and 3 women, and therefore the sequence of a specific chromosome tract accessible on the net actually belongs to only one of these individuals. It follows that the reference sequence of each human chromosome found in the network is a ***patchwork*** (sewing work in which pieces of fabric of different provenance are joined) in which each chromosomal segment comes at random from one of the 8 different anonymous individuals. Venter used a similar strategy starting with 5 individuals.

The two approaches were also different technically. The consortium used a ***traditional approach***: the pieces of DNA are similar to fragments of the pages of a book; each book, which contains a certain number, is comparable to a chromosome and the set of volumes / chromosomes constitutes the encyclopedia / genome. First the position of the isolated DNA segment in the test tube was located, thus identifying that it was in that book and on that page, marking it with fluorescent dyes and observing the link to a certain chromosome under the microscope. The segment thus mapped was further broken up into fragments with each of them sequenced and the sequences reassembled in a single sequence until the chromosome is complete. Since the fragments are generated by random breaks, their sequences are partially superimposed: we imagine cutting up a page and then having to recompose it - it would be very helpful if we had another equal page cut in a different way, where the last words of a fragment of the first page are with the first of the new one.

Venter used an alternative ***global approach***: he didn't bother to know where the DNA fragment he sequenced was positioned, but he finely chopped the DNA with ultrasound, determined the base sequence of each fragment and with a very powerful bioinformatic analysis determined the continuity of the fragments derived from each chromosome assembling them in a single sequence. However, Venter's assembly had some additional shortcomings. Data for anchoring specific chromosomes were missing, data that the other project had instead, which took more than 10 years while Venter's group finished the study in 3 years. Some have noted that *Celera Genomics* also had information that was gradually released by the public consortium, a circumstance that could facilitate the positioning of segments along the chromosome.

The sequencing of the human genome cost 1 dollar per base, thus 3 billion dollars in total. Since the mid-2000s, new sequencing methods have brought down the cost of DNA sequencing, suggesting that the time when a human genome can be entirely sequenced in 1 day at a cost of 1,000 dollar is not far off. Thanks to these methods, projects have been launched to sequence thousands or tens of thousands of human genomes to identify sequence

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

variants related to specific characteristics and in particular to predispositions to disease (the most ambitious is now the English *100,000 Genomes Project*, http://www.genomicsengland.co.uk/).

Types of variability in the human genome

At the molecular level the differences between the genomes of different human individuals are of ***three types*** (Figure 3.6). These differences concern any two human individuals, while it has not been possible to substantiate the concept of distinct human "races" at the genomic sequence level.

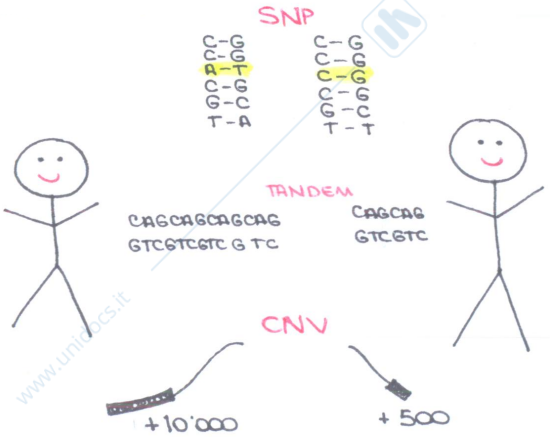


Figure 3.6. Types of variability in the human genome. Illustration by Amalia Sanna Passino.

A first source of variation is due to the ***SNPs*** (pronounced "snip"), ***single nucleotide polymorphisms***: they are precisely those simple variants for which an individual has a base and the other has a different one in a certain position. On average ***about 1 nucleotide out of 1,000*** is polymorphic, therefore we can conclude that from this point of view, any two human genomes are ***99.9% equal***.

Tandem repetitive sequences make it possible to differentiate two individuals as the number of repetitions of the basic module can vary. For example, at a certain point in the genome presenting one of these repetitions, an individual could present the CAGCAG sequence, namely $CAG_{(2)}$, in which the CAG module is repeated twice, and another $CAG_{(13)}$, in which the repetition of the basic unit is carried out 13 times. In some sites the number of repetitions can be extremely polymorphic, and by studying a certain number of such sites it is possible to determine a molecular "signature" virtually specific to that individual, being extremely unlikely (although not formally impossible) that another individual presents exactly the exact same number of repetitions in several places.

In 2004 a third class of variations was described: the s***ubmicroscopic structural variants*** (review in Feuk et al., 2006). They are ***variants of the chromosome*** which, as the name implies, are much more extensive than a few nucleotides, arriving to modify the structure of the chromosome, although not to the point of being visible under the optical microscope.

The most common variation consists in the presence or absence of segments whose size can range ***from*** around 1***,000 bp (1 kbp) to 3 Mbp***. Consequently the number of copies of these segments can vary from one human chromosome to another because the segment can be gained (***gain***) or lost (***loss***), hence the name of ***Copy Number Variants*** or ***CNV***. For example, there may be a 10 kbp segment in one chromosome present in two copies in the genome in a certain individual and completely absent in the genome of another. We can deduce that, given the size of the segments involved, the percentage difference in sequence between individuals, which we had seen to be equal to 0.1%, may actually reach 0.5% or more. However, molecular biology experiments are required to identify CNVs since the chromosome would appear to be normal under the microscope. A recent article published in Nature (Sudmant et al., 2015) presented the distribution of CNVs in 2,504 human genomes. It is not always clear whether a CNV constitutes a normal variant of the human genome or is associated with a disease or a predisposition to it. One of the most original results of this study is the observation that some knock out genes, therefore absent or incomplete homozygous genes (for which both copies of the gene are not functional), occur in nature in normal subjects: this suggests that we can do without a number of human genes. This underlines the complexity of the interactions between genes from the point of view of the overall genome function.
Other submicroscopic variants do not involve a difference in the number of copies of a sequence but rather its orientation (***inversions***).

Due to ***somatic mutations***, even the cells of the same individual will have a percentage of genome variability. Considering the SNPs, at least one base pair on a billion ($1/10^9$) changes at every cell replication. At each mitosis, therefore, 6-7 point mutations occur, so it can be estimated that an adult cell, separated by at least 46 cell divisions from the zygote, has accumulated about 300 mutations. They normally go unnoticed because, being random, they are unlikely to affect a functional element of a gene, and moreover this should occur in a gene expressed in the cell type affected by the mutation and important for the normal function of that cell. It can be concluded that in the human organism no cell is exactly the same as another. If we also consider that we can obtain the complete genomic DNA sequence only by irreparably damaging the cell that contains it, the formal demonstration results that we cannot know the complete DNA sequence of a living cell (Strippoli et al., 2005).

Genome variability among different species

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

The variability of the genome is greater, as expected, if we consider two individuals of different species. In the case of the biologically most similar species to *Homo sapiens*, the chimpanzee (*Pan troglodytes*), the base pairs are 98.8% similar (in this case the difference is slightly higher than 1%) (Chimpanzee Sequencing and Analysis Consortium, 2005). However, this is sufficient to make almost all proteins different even if they are just a few amino acids. Between man and mouse 85% of similarity is calculated, and even with bacteria 20% similarity is found, due to the preservation of genes fundamental for the basic biological mechanisms common to all living beings.

Allele concept

In the initially published map of the human genome, the sequence of a single illustrative chromosome was determined and represented. However, in the human genetic set there are two copies of each non-sexual chromosome (autosome), so for each type of chromosome there is a sequence of maternal origin and one of paternal origin. This means that all the genes that are located on an autosome are present in duplicate, and since the mother and father are two different individuals, it is evident that they too will have the sequence variability that characterizes the human genome. Therefore a certain gene could present in a genome with two different sequences, for example in a certain position an individual could have the base C in the sequence of maternal origin and the base A in that of paternal origin - in practice an SNP, for which this gene has two variants. This is the minimum variation that can differentiate the two genes: a single base, and assuming that the opposite one is complementary to it, a pair of bases. A pair of bases is enough to distinguish one allele from another, a variant of a gene from another variant form in which it can present itself.

The ***allele*** can therefore be defined as an ***alternative form of a gene***, an alternative sequence with which a gene can occur, which could also be called a gene variant.

If the variation is inside a coding sequence, which can happen in about half of the cases, two different proteins could be formed, since an amino acid sequence is produced by reading the maternal mold and the other by reading the paternal one. Thus the organism will be characterized by a mixture of these two proteins, possibly different even for just one amino acid, but still not structurally and possibly even functionally identical.

There are two possibilities: if the two parents have different variants of a gene, we are different from the same gene, that is ***heterozygous*** where there are ***two different copies of the gene*** in the same cell (in the same cellular genome).

If instead the two ***sequences*** are completely ***identical***, then that individual would be ***homozygous*** for that gene.

It is evident that an individual can be heterozygous for a gene and may be homozygous for the gene immediately after, so *for each gene we must determine whether the individual has the same or different sequences*.

The last human genome sequencing projects evaluated this parameter, while in the first studies only one example chromosome of the two actually present was sequenced. In 2007 the first sequence of a ***human diploid genome*** was published (sequence "HuRef", obtained from the DNA of Dr. Craig Venter cited above; Levy et al., 2007). It was thus possible to demonstrate that the ***percentage of heterozygosity***, i.e. how many paternal genes occur with a different sequence than the maternal ones, can reach about half of the genes (***44%***). In 17% of the total human genes the SNP that causes heterozygosity can be found in the portion of the gene encoding for proteins and be "not synonymous", that is leading to the production of ***two different proteins*** for at least one amino acid. Thus we can have a significant number of genes (around 4,000) that produce two slightly different proteins in the same organism. This is nothing more than the molecular basis of all ***Mendelian genetics***: the two variants of the genetic determinant are separated in the gametes of the two parents and are distributed randomly in the children.

3.5 Human genes (number, structure and function)

A human medium **mRNA** consists of about **3,500 nucleotides**. This data has a meaning, since this value is included within the typical dimensions of a sequence that can be inserted into a vector, and that therefore can be proposed for gene therapy. When dealing with a genetic disease, it is very important to know if the gene's mRNA that is the cause is large or small, because if it is average, you can still think of replacing it with a vector, but certain large dimensions are absolutely incompatible with vectors that can be used for gene therapy.

Similarly to say that a gene is large or small can change a lot compared to the diagnosis or the study of the gene, because in sequencing, except for the introduction of more powerful methods of limited use, the times are very different: if the gene is small it can be sequenced even in a day (for example 1,000 bases), while if it is one of the largest genes (about 2.5 Mbp), sequencing it in detail would take months.

These are only examples to show how the main statistical data on the size and structure of human genes (summarized in Figure 3.7), in addition to being the object of study to understand if there is a biological logic in these values, also lead to a series of practical consequences.

Next we will discuss the class of ***protein-coding genes*** first, then that of ***non-coding genes***.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

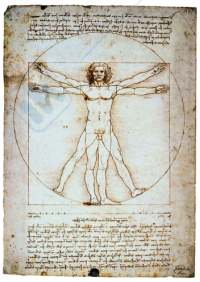
www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Protein-coding genes - Number

There are about ***20,000 protein-coding genes*** in humans, but we will never know how many we can be sure of. This is because, while it can be affirmed with certainty that a gene exists because a genetic product (RNA) derives from it, it cannot be absolutely affirmed that a sequence is *not* a gene only because we do not trace the corresponding product. In fact, the gene could be transcribed only in a spatially and temporally restricted way, for example only in the cerebellum of a newborn up to 6 months, and until such a type of sample is not studied the data will not be known. Furthermore, a gene could be activated only in certain pathological circumstances, or in response to particular environmental stimuli or the administration of a drug.

Drosophila melanogaster
~14.000

Caenorhabditis elegans
~21.000



Homo sapiens
~22.000

The approximate total number of human genes as emerged from the completion of the Human Genome Project was a surprise because man was expected to have many more, considering that the fruit fly has about 14,000 and the very simple worm *Caenorhabditis elegans*, composed of only 959 cells, has about 21,000 genes.

Therefore the quantity of genes present does not seem to be a correct and reliable indicator of the complexity of an organism, at least as regards the genes encoding proteins, i.e. genes that produce mRNA.

The number of genes annotated in the public databases in 2019 is ***19,116*** but this number is certainly at fault because a certain number of genes has yet to be characterized in detail and does not figure in the databases, but these genes certainly exist. *A human chromosome therefore contains* **797** *known genes on average.*

Protein-coding genes - Dimensions

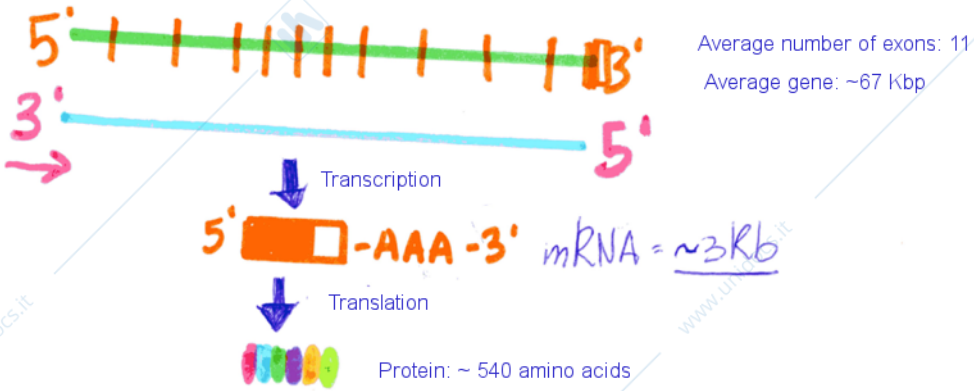


Figure 3.7. Scheme of a "medium" human gene. Drawing by Anna Mingazzini, 2016.

The ***average size of a human gene*** is about ***67 kbp*** (67,000 base pairs) (Figures 3.7 and 3.8; data from Piovesan et al., 2016, and Piovesan et al., 2019, manuscript sent for publication). Compared to this average value, a gene of 1,000 base pairs can be considered small while a gene extended to 1 million base pairs is very large. In fact the quantitative parameters of human genes are subject to an extraordinary ***variability***, considering that the gene for a smaller protein, currently filed, consists of about 200 base pairs and produces a very small polypeptide, while a larger gene is no longer considered the one for dystrophin (the gene for the mutated protein in muscular dystrophy, extended for more than 2 million base pairs), but rather a gene of about 2 and a half million base pairs (2.47 Mbp), *RBFOX1*.

The ***average size of a human mRNA*** is about ***3.5 kbp*** (3,500 base pairs).

These two fundamental numbers make us immediately understand that a messenger is much smaller than the gene, that is that the introns, the sequences removed by the messenger that we will therefore no longer find in the mature messenger, represent in fact most of a human gene: only 3.5 kbp of sequence, of the original 67 present on the DNA, remain in the mRNA.

The protein-coding sequence (coding sequence, CDS, which also includes the stop codon) in turn is about half the length of a messenger because ***messenger RNA also consists of untranslated sequences*** (5′UTR, 3′UTR): only the sequences present between the start codon and the stop codon are translated.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

All this must lead to the cancellation of the typical mental scheme according to which the gene consists of an alternation of introns and exons, all with similar dimensions, and to the acceptance of the image corresponding to the truth that *exons are small sequence islands that emerge occasionally from an ocean of introns*.

HUMAN GENOME	[2001-2019]
Aploid size=**3.2 Gb**	
CHROMOSOMES	22 (×2) Autosomes
Largest: 1 (249 Mb)	Smallest: 21 (47 Mb)
sex chromosomes: X (156 Mb) – Y (57 Mb)	
mtDNA (16,569 bp)	
32% of DNA: PROTEIN-CODING GENES	
About 20-25,000	Known: **19,116** [2019]
Mean: 797/chromosome	(35/band 550 – 1/167 kb)
Gene	Mean size=**67 kb**
0.2 kb (*KRTAP6-2*) to 2,474 kb (*RBFOX1*)	
Mean intergenic distance: 100 kb	
mRNA	Mean size=**3.5 kb**
0.3 kb 5´UTR + 1.7 kb CDS (552 codons) + 1.6 kb 3´UTR	
Exons	Number: 1 to 363 (*TTN*); Mean=11
	Size: Mean 311 bp (2 to 27,303 bp)
	Total: 5% of gene sequences
Introns	Number: 0 to 362 (*TTN*); Mean=10
	Size: Mean 6.9 kb (30 to 1,160,411 bp)
rRNA	28S=5.070 Kb – 18S=1.869 Kb
	NR_003287 NR_003286

Figure 3.8. Summary data on the human genome (2019).

The average protein, therefore, is obtained knowing that a codon of three nucleotides encodes an amino acid: if the CDS is 1.7 kbp on average (1.655 nucleotides) and we are going to divide it by 3 we obtain that the ***average size of a human protein*** *is* ***551 amino acids*** (Figure 3.7), not considering the stop codon. Thus a polypeptide of a few amino acid tens can be considered small, while a protein composed of thousands of amino acids is large.

Protein-coding genes - Structure

The data summarized in Figures 3.7 and 3.8 are briefly commented on here (Piovesan et al. 2016; Piovesan et al. 2019).

On average a human gene has 11 exons, or 11 traits of DNA that are kept in the messenger, but even in this case the variability is enormous. In fact a gene can have ***only one exon (monoesonic genes)***, which means that it is not subjected to splicing, since to have an intron it must be composed of at least two exons, while at the other end we have the gene for the titin muscle protein which has the ***maximum number of exons*** described, ***363***.

The ***average number of introns*** in a human gene is ***10***. In fact, being in the midst of exons, the number of introns in a gene is always equivalent to the number of exons minus one. The number of introns can therefore go from a ***minimum of 0***, in monoesonic genes, to a ***maximum of 362***, in the case of the gene that has 363 exons.

Exons are much shorter than introns with the ***average size of a human exon*** about ***300 bp*** and that of a ***human intron*** about ***6.900 bp***. An intron therefore turns out to be on average 22 times more than an exon.

However, the length of an exon can exceed 27,000 bp (27 kbp). The ***minimum size*** of an ***exon*** has no constraint. In fact we can have in sequence an intron, then a very small exon, virtually also constituted by a single pair of bases, and then the next intron. Once the two introns in the messenger are removed, the exon will remain, whatever its size. In reality, the ***minimum exon*** is ***2 bp***. As for the ***intron***, there are limits: it is impossible, for example, that an intron is less than 4 bp long because it must have the GT sequence at the beginning and AG at the end in order to be subjected to splicing, but in reality it must also have a poly-pyrimidine trait (a trait rich in pyrimidines) near the "splicing acceptor" site i.e. AG. Since this trait does not have a fixed length, then the theory will not be given a clear definition of how long an intron can be at a minimum. One of our recent bioinformatics studies has defined the ***minimum length of a human intron*** of ***30 bp*** (Piovesan et al., 2015). This ***limit*** is actually ***valid for all species*** studied so far. Determining the sequence of really existing minimal introns allows us to understand which sequence elements are essential for splicing.

The ***longest*** described ***intron*** exceeds 1 million base pairs (***1 Mbp***): this shows that there are single introns enormously longer than the average length of an entire gene.

We note in passing that determining the maximum and minimum limits of a biological parameter by recording them in a sort of "Guinness Book of Records" (where primates are records ... not great apes) is not a statistical curiosity, but as shown in the example of minimal intron provides clues about the operating constraints of a certain biological mechanism. Another example is the correlation between telomere length and life expectancy: the theory of progressive telomere shortening predicts that a man cannot live for more than about 120 years. In the "Guinness Book of Records" it is recorded that the longest certified age ever reached by a member of our species is 122 years.

Why does the cell spend an enormous amount of energy transcribing a very long segment into RNA, even a million base pairs, and then cut this intronic sequence from the transcript? It is not possible for this to happen by chance because the cell is very efficient in energy management and would not waste all this energy to transcribe a segment only to then destroy it. The answer to this question is emerging with the demonstration that these ***introns*** contain many ***functional*** sequences, i.e. some traits of the introns are excised and function as non-protein-coding regulatory and RNA sequences that modulate the function of other genes. Furthermore, an abundance and complexity of the sequences that can function as introns is at the base of the alternative splicing and therefore of the possibility of obtaining different products, interacting in a complex way, starting from a single gene.

This observation leads us to a basic concept that occurs in Biology and Medicine, and generally in nature, that is ***if something exists it is useful***, but often we simply don't know what. In many classic Genetics texts there were definitions of introns such as "*junk DNA*" or "remnants of evolution". In reality, the study of mutations shows ever more clearly that the modification of non-coding sequences involves functional effects, just as it happens analogously if we remove any structure of our organism. To confirm this, a few years ago it was shown that the

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

appendix, used as a classic example of a rudimentary organ, residual of evolution and no longer functional, produces specific classes of antibodies that are drastically reduced in serum following appendectomy (Andreu Ballester et al., 2007). Moreover, the removal of the tonsils in later years favors a reaction in the appendix that increases the risk of acute appendicitis, already known in the 1940s (and described in the hospital novel "Corpi e anime") and only recently rediscovered (Andreu-Ballester et al., 2003).

Protein-coding genes - Functions

A very rough but useful subdivision allows to identify 5 large classes in which human genes can be grouped according to their main function.

22% of the genes in our genome encode for any of the enzymes that operate in all human ***metabolic pathways*** (so about 1/5 of human genes are committed to making metabolism enzymes);

25% of genes are involved in the ***management of genetic information*** itself, i.e. the genome uses 1/4 of its genes to maintain the structure and function of the genes themselves, so around 6,000 genes are involved in these processes: DNA replication; DNA repair; distribution of DNA molecules in cells (mitosis and meiosis); transcription; transcript processing (splicing, capping, polyadenylation); translation into protein;

in ***20%*** of the cases the genes are ***structural***, therefore committed to synthesize the various proteins that ensure the basic structure of the cells, such as membrane proteins, proteins of the inner membranes, of the cytoskeleton and so on;

12% of human genes are genes for ***signaling pathways***, which is a very large number compared to less complex organisms. This can be considered as one of the many possible keys to explain why the human genome, despite having few genes, is the basis of such great complexity. In fact, cellular signals, often proteins, connect each cell with other cells, exponentially increasing the complexity of the organism. The proteins that intervene in the signaling mechanism can be divided into signal proteins, able to send a signal by binding to other proteins called receptors, which will send the signal through other proteins called transduction, to the nucleus, where other genes will be specifically activated in response to this signal.

Many genes of these first 4 classes are active in most cell types, in fact all cells must somehow metabolize, and manage genetic information, have a structure and exchange signals.

20% of genes encode proteins that are not typically present in all tissues and are therefore expressed in a ***tissue-specific*** manner. In extreme cases a protein can be produced only and exclusively in a cell type, such as antibodies (B lymphocytes), hemoglobin (red blood cells), rhodopsin (retinal photoreceptors), insulin (beta cells of pancreatic islets) and so on.

Non-coding genes

1. rRNA

Ribosomal RNAs (rRNAs) are transcribed from a single transcription unit, so a single RNA is initially formed, which is then divided into RNAs 28S, 5.8S and 18S. This sequence is called the ***ribosomal RNA gene (rDNA)*** and is repeated in tandem 80 times at the level of the short arm of the ***acrocentric chromosomes*** that are 13, 14, 15, 21 and 22 in humans. L '5S rRNA does not derive from this structure, but from a gene present in 2,000 copies distributed in all chromosomes, in particular chromosome 1.

The rDNA sequences are in total 800, that is 80 units in 5 acrocentric chromosomes each of which is present in two copies, one derived from the father and another from the mother.

The nucleolus assembles itself around the rDNA, in the nucleus, so that the rRNAs gradually produced can be incorporated into the ribosomes. It follows that the maximum number of nucleoli that a cell can have is 10, one for each available acrocentric chromosome.

The genes for rRNA have a ***high gene dosage***, i.e. there are many molds for the production of the product. The gene dosage is the ***number of copies of a gene in a cell***. For example, the molds for insulin (gene on chromosome 15) are 2, one of maternal origin and the other of paternal origin, a number much lower than that of the genes for rRNA (800). The reasons for such a high dosage can be understood in light of the fact that rRNAs constitute the vast majority of cellular RNA, around 85%, while tRNAs are 10% and mRNAs together are only 1-5%. Moreover, such redundancy also ensures protection in the case of mutations, since the genes for rRNAs are indispensable for the survival and functioning of any cell.

2. tRNA

If we have 61 codons for amino acids we should have a ***transfer RNA (tRNA)*** to read each codon, but since the anticodon of the same tRNA can match different codons, there are only 48 types of tRNA (coded by a total of about 450 genes) in man which are sufficient to read all 61 codons of the genetic code (which includes 64 triplets in total but 3 constitute stop signals of the translation).

These genes are also redundant. On average, a tRNA is present in 10 copies in the human genome, and they are found scattered in all chromosomes except for chromosomes 22 and Y, while sometimes they can be found grouped in small "clusters", in particular on chromosomes 1, 6 and 7.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

3. ncRNA

This new class of RNAs that are not translated into protein (noncoding RNA, ***ncRNA***), and are not even rRNA or tRNA, is produced by sequences that are ***genes*** in all respects because they are endowed with sequences similar to ***promoters***, are transcribed in ***specific cell types***, and if they are mutated they give rise to ***pathologies*** (Mattick et al., 2006).
The best known classes are: spliceosomal RNA (***uRNA***), which are incorporated into splicing proteins; small nuclear RNAs (***snRNA***), also implicated in splicing; small nucleolar RNAs (***snoRNA***), which are found in the nucleolus and are incorporated by proteins that then catalyze a reworking of ribosomal RNAs, that is they modify the bases of rRNAs, adding for example pseudouridine.

In the last few years the class of microRNAs (***miRNAs***), small RNAs of 20-22 nucleotides, which regulate the genes for the expression of messengers, has acquired particular importance, for example lowering the level of translation of that messenger following the link with it. Thousands are described and many have yet to be characterized.

The small interfering RNAs (***siRNA***) work with a general mechanism similar to microRNAs, of which they are larger. They bind to mRNAs and regulate their expression, in particular triggering their degradation. The Nobel Prize for Physiology or Medicine was awarded in 2006 to Andrew Fire and Craig Mello for this discovery (Fire et al., 1998).

The genes for non-coding RNAs may be ***overlapped*** on each other or even ***nested*** inside each other, in both filaments, both in one filament and on the opposite one (Figure 3.9). They can also be located within genes for proteins, complicating the classical scenario that involved the presence of only one gene in each locus.

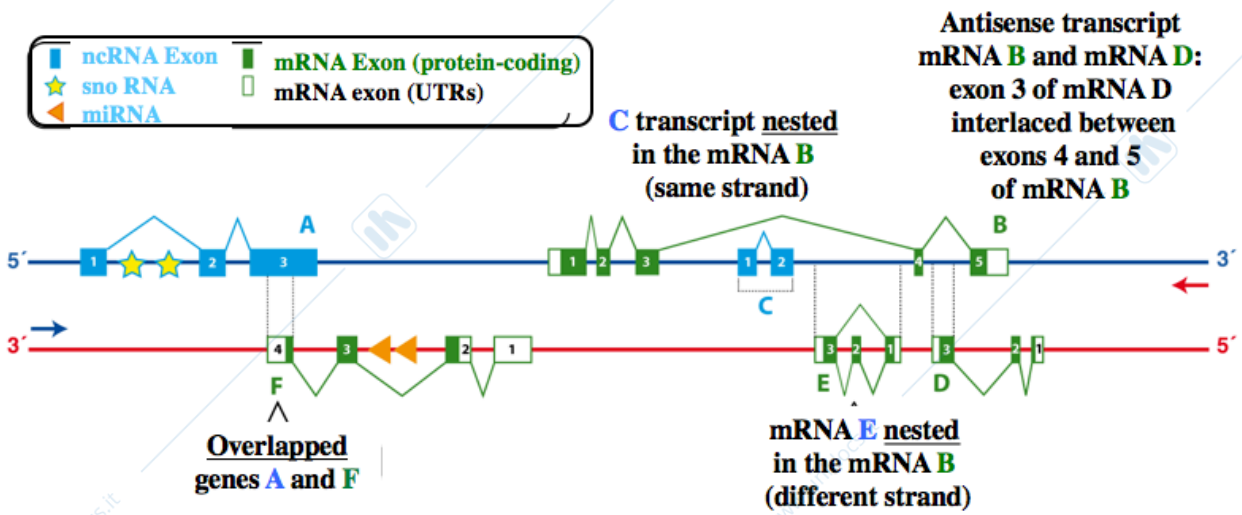


Figure 3.9. Complexity of the genomic organization of human genes.

The recent results of the ENCODE Project (http://genome.ucsc.edu/ENCODE/) have shown that due to the many types of non-protein coding genes, the ***percentage of transcribed DNA can reach 80% of the human genome*** ("pervasive genome transcription"), modifying the classical scenario that envisaged a small functional part interspersed with "intergenic" sequences.

The number of genes for ncRNA appears to increase with the complexity of organisms, and in humans their total number is now estimated at 20-25,000, equivalent to that for coding RNAs.

Web sites

The collection of links to the mentioned genomic sites and to others can be found online at:
http://apollo11.isto.unibo.it/appunti.htm

Note: It is possible to search for the summary and also the complete text of most of the articles cited by consulting the public database PubMed: http://www.ncbi.nlm.nih.gov/pubmed/
To access the complete text of publications in journals to which the University of Bologna is subscribed, follow these instructions:
https://sba.unibo.it/en/almare/almare-tools-and-services/remote-access

References

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Andreu-Ballester JC, Colomer RE, Millán Scheiding M, Ballester F. Tonsillectomy and risk of acute appendicitis. Am J Gastroenterol 98:2806-2807, 2003.
http://www.ncbi.nlm.nih.gov/pubmed/14687841

Andreu-Ballester JC, Pérez-Griera J, Ballester F, Colomer-Rubio E, Ortiz-Tarín I, Peñarroja Otero C. Secretory immunoglobulin A (sIgA) deficiency in serum of patients with GALTectomy (appendectomy and tonsillectomy). Clin Immunol 123:289-297, 2007.
http://www.ncbi.nlm.nih.gov/pubmed/17449327

Asimov I. *The genetic code*. New American Library of World Literature, New York, 1962. [Tr. it.: *Il codice genetico*. Einaudi, Torino, 1968].

Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, Vitale L, Pelleri MC, Tassani S, Piva F, Perez-Amodio S, Strippoli P, Canaider S. An estimation of the number of cells in the human body. *Ann Hum Biol* 40:463-471, 2013.
http://www.ncbi.nlm.nih.gov/pubmed/23829164

Campillo-Balderas, Lazcano A, Becerra A. Viral genome size distribution does not correlate with the antiquity of the host lineages. Front Ecol Evol, 2015.
http://journal.frontiersin.org/article/10.3389/fevo.2015.00143/full

Cavelier L, Johannisson A, Gyllensten U. Analysis of mtDNA copy number and composition of single mitochondrial particles using flow cytometry and PCR. *Exp Cell Res* 259:79-85, 2000.
http://www.ncbi.nlm.nih.gov/pubmed/10942580

Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69-87, 2005.
http://www.ncbi.nlm.nih.gov/pubmed/16136131

Ciccodicola A, D'Urso M. *Il sequenziamento del genoma umano*. In: "Le Scienze quaderni" n. 100/D: *Il progetto genoma*, a cura di Dulbecco R, Vezzoni P. Milano, 1998, pp. 29-33.

Eddy SR. The C-value paradox, junk DNA and ENCODE. Curr Biol 22:R898-899, 2012.
http://www.ncbi.nlm.nih.gov/pubmed/23137679

Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc Lond B Biol Sci 370:20140331, 2015.
http://www.ncbi.nlm.nih.gov/pubmed/26323762

Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet* 7:85-97, 2006.
http://www.ncbi.nlm.nih.gov/pubmed/16418744

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 391:806-811, 1998.
http://www.ncbi.nlm.nih.gov/pubmed/9486653

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 409:860-921, 2001. [Questo articolo scientifico è disponibile liberamente su Internet all'indirizzo: http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html].

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 431:931-945, 2004.
http://www.ncbi.nlm.nih.gov/pubmed/11237011

Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, et al. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* 15:141-161, 2015.
http://www.ncbi.nlm.nih.gov/pubmed/25722247

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, Macdonald JR et al. The Diploid Genome Sequence of an Individual Human. PLoS Biol 5:e254, 2007.
http://www.ncbi.nlm.nih.gov/pubmed/17803354

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Mattick JS, Igor V. Makunin IV. Non-coding RNA. *Hum Mol Genet* 15:R17-R29, 2006.
http://www.ncbi.nlm.nih.gov/pubmed/16651366

Mohanta TK, Bae H. The diversity of fungal genome. *Biol Proced Online* 17:8, 2015.
http://www.ncbi.nlm.nih.gov/pubmed/25866485

Piovesan A, Caracausi M, Ricci M, Strippoli P, Vitale L, Pelleri MC. Identification of minimal eukaryotic introns through GeneBase, a user-friendly tool for parsing the NCBI Gene databank. *DNA Res* 22:495-503, 2015.
http://www.ncbi.nlm.nih.gov/pubmed/26581719

Piovesan A, Caracausi M, Antonaros F, Pelleri MC, Vitale L. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database (Oxford)*, 2016. doi: 10.1093/database/baw153
https://www.ncbi.nlm.nih.gov/pubmed/28025344

Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol 14:e1002533, 2016.
https://www.ncbi.nlm.nih.gov/pubmed/27541692

Strachan T, Read AP. Organization of the human genome. Capitolo 9 in: Strachan T, Read A: *Human molecular genetics 3* (3ª edizione), Garland, New York, 2004, pp. 240-274. [Tr. it..: *Genetica umana molecolare*, UTET, Torino, 2006, pp. 227-259].

Strippoli P, Canaider S, Noferini F, D'Addabbo P, Vitale L, Facchin F, Lenzi L, Casadei R, Carinci P, Zannotti M, Frabetti F. Uncertainty principle of genetic information in a living cell. *Theor Biol Med Model* 2:40, 2005.
http://www.ncbi.nlm.nih.gov/pubmed/16197549

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. An integrated map of structural variation in 2,504 human genomes. *Nature* 526:75–81, 2015.
http://www.ncbi.nlm.nih.gov/pubmed/26432246

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. The sequence of the human genome. *Science* 291:1304-1351, 2001.
http://www.ncbi.nlm.nih.gov/pubmed/11181995

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

4. GENE FAMILIES

Definition

Families of DNA sequences are made up of ***groups of similar sequences***. The similarity between two DNA sequences (technically "similarity") can be objectively quantified, as in this example:

Sequence 1 → CCATT
Sequence 2 → CCATG
Sequence 3 → CCCCC

Sequences 1 and 2 are similar for 80% since we are able to align 4 bases out of 5, sequence 3 is instead 40% similar to the first. Two sequences could also have no coincidence point, and be 0% similar. If we extend this reasoning to the actual gene sequences we discover that some genes are very similar to each other and produce proteins whose amino acid sequence will turn mutually similar.

In molecular genetics the concept of similarity is formally distinguished from that of homology. The ***similarity*** simply indicates that the two sequences are similar. The ***homology*** implies that the two sequences derive from a ***common ancestor***. Thus homology always implies a similarity: in fact if the two sequences have a common ancestor they are necessarily similar to each other. On the contrary, similarity does not necessarily imply a homology because it cannot be formally demonstrated that the two sequences had a common ancestor even millions of years ago or they might not have had it and be similar by chance or by progressive convergence towards the same sequence due to mutations.

We distinguish families of gene sequences, that is, ***gene families***, if the sequences are similar to one another, from the ***nongenic sequence families*** in which the sequences that make up the family are not active genes.

Gene families

They are similar DNA sequences, arbitrarily defined as more than 70% identical for nucleotide positions, located in different points of the genome and transcriptionally active. In other words, they are different genes, located in different loci but similar to each other.

We distinguish three types of genetic families.

1. In classical gene families, similarity is maintained throughout the gene extension.

These families are the most typical ones. The classic example is that of the family of oxygen carrying globins in which the gene for the alpha globin is similar to that for beta globin, but also to that for the gamma globins, deltas etc. Other typical families are histone genes (in fact histone proteins are very similar to each other and are encoded by similar genes), the rRNA family (transcriptional units repeated several times on the same chromosome and in more chromosomes and have the same sequence), the family of genes for cytokines but also that of genes for cytokine receptors, the family of genes for myosin, keratins and so on. In the international nomenclature these genes are indicated with the same name followed by a progressive number, often assigned according to the moment in which they were identified, for example keratin 8 (*KRT8*), keratin 20 (*KRT20*) and so on up to keratin 40 (*KRT40*). We recall that the official symbols of each human gene are defined internationally and are written in italic capitals, reserving the round character (not italic) for the corresponding proteins.

To constitute a family it is necessary that at least two components of the same family exist, however there are also very numerous genetic families. The gene family that accounts for the greatest number of components in the human genome is the family of olfactory receptor genes, which has about 950 members, distributed on almost all chromosomes and coding for membrane proteins exposed on the cilia membrane of olfactory neuronal cells. It is believed that each olfactory cell expresses a single receptor protein, specific to detect a specific odoriferous molecule, even if a recent study conducted in mice shows that up to 40% of mature olfactory neurons express more than one gene for olfactory receptors.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

It is intuitive that if some genes for proteins are similar to each other, probably the proteins they codify will also be, and so we pass to the concept of ***family of proteins***. Observing the alignment of several amino acid sequences belonging to proteins of the same gene family (Figure 4.1) we see that in some points the amino acids are different, while in a stretch we find an "array" of 9 aligned amino acids of which 8 are exactly identical in all the proteins considered. The conservation in a certain position of a constant sequence composed of the same amino acids suggests that this specific sequence plays a fundamental role for the function of all the proteins that contain it.

Figure 4.1 Amino acid sequence alignment of two traits of DSCR1-Like gene proteins in different species (prefix m = mouse; DROME=*D. melanogaster,* fruit fly; CAEEL=*C. elegans,* nematode worm; SCHPO=*Saccharomyces pombe* and YEAST=*Saccharomyces cerevisiae,* yeasts; Adapt78 is a hamster protein, DCSCR1L2, ZAKI-4 and DSCR1 are human proteins, with DSCR1.1 and DSCR1.4 two forms of DSCR1 derived from alternative splicing of the *DSCR1* gene are indicated, a phenomenon that in the traits shown does not lead to different amino acid sequences). Alignment with ClustalW of sequences present in public databases.

The alignment taken as an example also shows that some amino acid sequences belong to the same species (*H. sapiens*), others to different species. The concept of ***gene and protein family*** thus extends to the similarity between sequences belonging to different organisms, suggesting that these sequences perform similar functions in all these organisms. It is therefore necessary to distinguish two phenomena: with ***paralogy*** we mean the presence of ***genes similar to each other in the genome of the same species***; for example the gene for human alpha globin is a paralog of the human beta globin gene, or, in Figure 4.1, *DSCR1L2* (now renamed *RCAN3*) is a paralogue of *ZAKI-4* e *DSCR1* (now renamed *RCAN2* and *RCAN1*, respectively), being that these genes are all human.

By ***orthology*** we mean the conservation of sequences of ***genes coding for proteins that perform the same function*** in ***genomes of different species***; for example, the human alpha-globin gene is an ortholog of the mouse alpha-globin gene, or, in Figure 4.1, human *DSCR1/RCAN1* is an ortholog of yeast *DSCR1L*.

When a gene family includes more members in the same species and corresponding genes in different species, the similarity between the respective orthologs (both genes and proteins) is higher than that between the respective paralogs; that is, for example, human beta globin will be more similar to beta globin than to human alpha globin.

2. In ***families of genes similar only in one region (domain)*** the sequences of the genes of the family appear similar only in a delimited portion called a domain.

The ***domain***, according to a classical definition, is considered as a ***stretch*** of protein sequence extended for at least 50 amino acids and ***that folds in a characteristic way***, assuming a constant conformation correlated to the performance of a characteristic function preserved in all the proteins that possess this domain. The rest of the sequence does not present a particular similarity.

For example, ***homeobox genes*** code for proteins that are transcription factors and are able to bind to specific DNA sequences to perform this function. In this case the alpha helix region is the one that gives the general ability to bind to DNA, adapting to the main groove, while other regions of homeobox proteins recognize specific

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

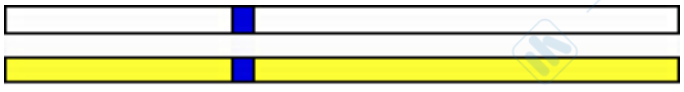
www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

sequences determining which promoters that particular homeobox protein can bind to. Therefore, aligning the sequences of human homeobox proteins a "block" of 60 amino acids will clearly emerge, called "box" of the homeotic or homeobox genes (Figure 4.2), in which the amino acids are very similar to each other, while for the rest of the extension of proteins one can have a very different sequence among the various members of the homeobox family.

```
HOXA1    1  MDNARMNSFLEYPILSS.....GDSGTCSARAYPSDHRITTFQ.SCAVSA
HOXB1    1  MDYNRMNSFLEYPLCNR.....GPS......AYSAHSAPTSFPPSSAQAV
HOXD1    1  .....MSSYLEYVSCSSSGGVGGDVLSLAPKFCRSDARPVALQPAFPLGN
HOXC4    1  ...MIMSSYLMD.........................SNYIDPKFPPCEEYSQ

HOXA1   45  NSCGGDDRFLVGRGVQIGSPHHHHHHHHHHPQPATYQTSGNLGVSYSHSS
HOXB1   40  DSYASEGRYGGG....LSSPAFQQNSGYPAQQPPS.....TLGVPFP.SS
HOXD1   46  GDGAFVSCLPLAAARPSPSPPAAPARPSVPPPAAPQYAQCTLEGAYEPGA
HOXC4   26  NSYIPEHSPEYYGRTRESGFQHHHQELYPPPPPRP.......SYPERQYS

HOXA1   95  CGPSYGSQNFSAPYSPYALNQEADVSG....GYPQCAPAVYSGNLSSPMV
HOXB1   80  APSGYAPAACSPSYGPSQYYPLGQSEG....DGGYFHPSSYGAQLGG...
HOXD1   96  APAAAAGGADYGFLGSGPAYDFPGVLGRAADDGGSHVHYATSAVFSGGGS
HOXC4   69  CTSLQGPGNSRGHGPAQAGHHHPEKSQ......SLCEPAPLSG.......

HOXA1  141  QHHHHHQGYAGGAVGSPQY..IHHSYGQEHQSLALATYNNSLSPLHASHQ
HOXB1  122  ....LSDGYGAGGAGPGPYPPQHPPYGNEQTASFAPAYADLLS.....ED
HOXD1  146  FLLSGQVDYAAFGE.PGPFPACLKASADGHPGAFQTASPAPGT......Y
HOXC4  105  ........................ASASPSPAPPACSQPAP.......

HOXA1  189  EACRSPASETSSPAQTFDWMKVKRNPPKTGKVGEYGYLGQPNAVRTNFTT
HOXB1  164  KETPCPSEPNTPTARTFDWMKVKRNPPKTAKVSEPG.LGSPSGLRTNFTT
HOXD1  189  PKSVSPASGLPAAFSTFEWMKVKRNASKKGKLAEYGAASPSSAIRTNFST
HOXC4  122  ...DHPSSAASKQPIVYPWMKKIHVS....TVNPNYNGGEPKRSRTAYTR

HOXA1  239  KQLTELEKEFHFNKYLTRARRVEIAASLQLNETQVKIWFQNRRMKQKKRE
HOXB1  213  RQLTELEKEFHFNKYLSRARRVEIAATLELNETQVKIWFQNRRMKQKKRE
HOXD1  239  KQLTELEKEFHFNKYLTRARRIEIANCLHLNDTQVKIWFQNRRMKQKKRE
HOXC4  166  QQVLELEKEFHYNRYLTRRRRIEIAHSLCLSERQIKIWFQNRRMKWKKDH

HOXA1  289  KEGLLPISPATPP..GNDEKAEESSEKSSSSPCVPSPGSSTSDTLTTSH
HOXB1  263  REEGR.VPPAPPG..CPKEAAGDASDQSTCTSPEASPSSVTS.......
HOXD1  289  REGLLATAIPVAP..LQLPLSGTTPTKFIKNPGSPSQSQEPS.......
HOXC4  216  RLPNTKVRSAPPAGAAPSTLSAATPGTSEDHSQSATPPEQQRAEDITRL
```

Figure 4.2 Amino acid sequence alignment of four of the many family members of the Homeobox human genes. The homeobox domain is highlighted in the yellow box. Alignment with ClustalW of sequences present in public databases.

3. In ***families of coding genes similar only for a small motif*** the similarity is limited to a small sequence composed of a few amino acids that are preserved, while all the others are different.



A classic example is the RNA helicase. For some reason all the enzymes capable of carrying out RNA molecules always have a sequence with the 4 amino acids: Aspartic Acid, Glutamic Acid, Alanine and Aspartic Acid (one letter code abbreviation: DEAD, hence the name "DEAD box" given to the motif). This small amino acid sequence is common to all helicases and is not usually found in other proteins.

Distribution of genes of the same gene family

As regards the ***localization of the genes that make up a gene family***, in the human genome they are usually dispersed on different chromosomes apparently at random and are therefore not found, as in bacteria, one after the other, although sometimes they are concentrated in groups called "clusters" (an example is represented by the genes for globins but also by the same homeobox genes). This arrangement for which very similar sequences are in a

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

contiguous position suggests that they derive from common ancestors and has been attributed to errors in DNA duplication or to unequal *crossing over* phenomena so that following an asymmetric exchange, resulting in an out-of-phase pairing of the homolog, a chromosome acquires a further copy of a gene which then changes over time, differing from the original copy.

Example: the globin cluster (Figure 4.3). On the short arm of chromosome 11 there are, relatively close to each other, the genes for beta-globin, delta-globin, gamma-globin and epsilon-globin, while genes for alpha- globin are found on another chromosome, although there is a strong similarity between the sequences of these genes and between the sequences of the corresponding proteins. Normally in hemoglobin, two alpha-globin chains suitably combine with two beta-, gamma- or delta-globin chains, therefore despite the similarities they have different roles. As for the reason for the existence of genetic families, it has been hypothesized that they originated from unique ancestral genes, that is that in very ancient organisms there was only one element and over the millennia an evolutionary process took place to ensure that with duplications of chromosomal segments (such as the one hypothesized for globins about 500 million years ago), with translocations (which would have brought the "ancient" alpha element onto another chromosome) or with gene duplications within the same chromosome, being that they obtained two elements that gradually acquired a slightly different functional value to the point that they now cooperate to make a protein complex that can have certain functions. The elements present on chromosome 11 give rise to proteins that have a different affinity for oxygen and therefore can intervene at different times in the development of the individual, for example a fetal hemoglobin is known to consist of alpha and gamma chains, while the adult is composed largely of alpha and beta. Representing the genes in this region of chromosome 11 according to the classical representation with lines for introns and rectangles for exons (by coloring in green the part encoding for protein and in white the untranslated regions), we can observe that the structures of the genes of Clusters are similar, with the coding sequence starting in exon 1 and ending in exon 3. The distances between the exons are of the order of 1 kbp, however in the genome these are relatively short distances. Analyzing the sequence, we find one that, as far as the DNA structure is concerned, maintains the exon and intron scheme, but it no longer has a usable coding sequence. This is called a pseudogene, and it is probably the result of a duplication of these elements, but succession of mutations accumulates and therefore these have altered the coding sequence for which it is no longer functional.

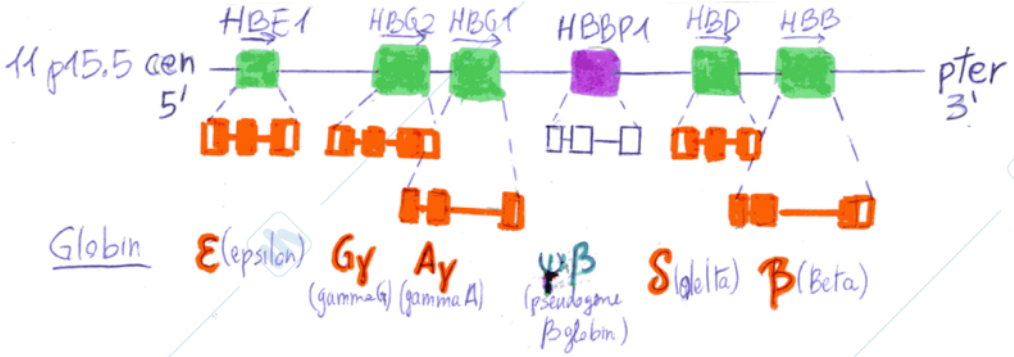


Figure 4.3. The cluster of globins on chromosome 11. Drawing by Anna Minguzzini

The homeobox genes are used to establish the formation of the different parts of the body of the animals and are preserved from the fruit fly to the man. A curious and unexplained fact is that the order of the homeoboxes on the genome reflects the order, from head to tail, of the parts of the animals; in fact there is no "molecular" reason why the gene that triggers the formation of the brain must be in the cluster, in the head, the one for the tail in the tail and so on.

From the ***functional*** point of view we sometimes observe the grouping of ***clusters of different genes*** not belonging to the same gene family but coding for proteins ***with similar functions*** or having a similar transcription modality: genes that must be transcribed at high intensity (in a way to exploit the localization in a region that has a high concentration of molecules of the enzyme RNA polymerase), housekeeping genes (constitutive, which therefore must always be active) and genes co-expressed in the same tissue, which, while coding for very different proteins, are activated simultaneously because their proteins have a similar expression profile by collaborating on typical and specific functions of that particular cell type.

Pseudogenes

Pseudogenes are gene-like sequences that have lost their function. Three types can be distinguished.

1. We speak of a ***solitary pseudogene*** when a single gene has lost its ability to function, presumably following the accumulation of mutations.

2. ***Nonprocessed*** or duplicated ***pseudogenes*** derive from DNA because they are similar to genes with introns and exons even if they are no longer expressed following mutation. There are at least about 11,000 in the human genome. Unlike disabled pseudogenes, which are genes present in a single copy which underwent inactivating mutations, these pseudogenes are degenerated copies of a gene that is still present and active in the genome, and for this reason they are also called "duplicates".

3. ***Processed*** or retrotransposed ***pseudogenes*** derive from mRNAs that have been retrotranscribed into DNA and inserted into the genome. There are about 8,000 in the human genome and due to their origin they resemble messengers (they have no introns). Also in this case, therefore, the gene that produced the mRNA from which the

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

pseudogene originated after retrotransposition is still present in the genome and active.

A significant proportion of the genome is therefore constituted by pseudogenes, which are assumed to no longer form RNA and are commonly considered residues of evolution. In reality, cases of ***transcribed pseudogenes*** are described and in this case their transcript may interfere with the messengers of the normal gene that has remained functional.

An interesting example of a disabled pseudogene is that for the hepatic enzyme gulonolactone (L-) oxidase which codes for the last stage of formation of L-Ascorbic acid (Vitamin C). The fact that we do not have this functional gene but rather in the form of a pseudogene means that we are completely dependent on the intake of Vitamin C from the outside from diet, unlike other animals that have the working gene. It has been discovered that on the short arm of chromosome 8 there is a structure that exactly follows the gene that in other mammals codes for the enzyme that catalyzes the last stage of the vitamin C synthesis path but has lost some exons, so the terminal part of the gene is missing. It is possible that the lack of this gene has been tolerated by man since he had a diet rich in fruits and vegetables. Several studies suggest that a chronic vitamin C deficiency may occur in the western diet, although light and therefore does not lead to obvious clinical manifestations of deficiency.

References

Buck L, Axel R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* 65:175-187, 1991.
http://www.ncbi.nlm.nih.gov/pubmed/1840504

Gerstein M, Zheng D. The Real Life of Pseudogenes. *Sci Am* 295:48-55, 2006.
http://papers.gersteinlab.org/e-print/sciam2/

Hanchate NK, Kondoh K, Lu Z, Kuang D, Ye X, Qiu X, Pachter L, Trapnell C, Buck LB. Single-cell transcriptomics reveals receptor transformations during olfactory neurogenesis. *Science* 350:1251-125, 2015.
http://www.ncbi.nlm.nih.gov/pubmed/26541607

Nishikimi M, Fukuyama R, Minoshima S, Shimizu N, Yagi K. Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. *J Biol Chem* 269:13685-13688, 1994.
http://www.ncbi.nlm.nih.gov/pubmed/8175804

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

NUCLEIC ACIDS EXTRACTION PRINCIPLES

MAIN EXTRACTION TECHNIQUES

Two categories:
1) Remove everything that is not DNA.
2) Selectively capture DNA.
Both classes of methods are used. They share two initial phases:
a) Cell separation.
b) Lysis of the cells.
Then, one of the two methods is applied to the lysate.

CELL SEPARATION

Aside from lymphocytes, where there are DNA rearrangements, DNA is virtually the same in all the somatic cells of an individual. Mitochondrial DNA can have slightly different sequences, and these can be distributed differently in each tissue. In tumours, it is different: cancer cells are estimated to have 50-100 deleterious mutations and can be part of subclones with different sequences. Moreover, they can be mixed with normal cells (e.g. stromal fibroblasts). In general, the most accessible tissue is peripheral blood, which also, being liquid, provides cells that are already isolated. Human saliva is even easier to be obtained, and it contains cells exfoliated from the buccal mucosa (epithelial cells) and lymphocytes. In this type of sample, however, the DNA is certainly more degraded than in the blood.

LYSIS

It is the phase in which the cells are broken, since DNA is complexed with proteins, and it is enclosed in two membranes.
The first goal is to put everything in suspension. This aim can be achieved with three fundamental methods:
a) Use of a **detergent**: it lowers the surface tension, binds to the lipids of the plasma membrane and on the other hand binds to the water by "tearing" the membrane. Generally, **SDS** (sodium dodecyl sulphate) is used.
b) Use of **enzymes** such as **proteases**. The best known is the protease K which, in small quantity, degrades proteins down to single amino acids.
c) Chaotropic agents ("agents able to induce chaos"): they disturb ordered structures, having at least one very large charged atom (ion) which at high concentration interferes with the three-dimensional architecture of proteins. Examples: sodium iodide (**NaI**), guanidine thiocyanate (**GuSCN**).
The strong denaturant activity of guanidine thiocyanate is also useful to inactivate endogenous RNases released during RNA extraction, in addition to its ability to destroy cells and solubilize cell components. Guanidine thiocyanate homogenate is obtained in the first step of the most popular method RNA purification from cells, described by Chomczynski and Sacchi in 1987.
Mechanical lysis may also be required in advance, e.g. in the case of solid tumours of a

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

very hard, wooden consistency (carcinomas).

DNA RECOVERY FROM THE LYSATE

1) EXTRACTION WITH ORGANIC SOLVENTS (classic method): phenol-chloroform. This method consists of several stages.

a) Extraction. It can be considered as a real "ex-traction", we separate the DNA from other undesired cell components; most of these are proteins. Organic solvents are generally apolar. DNA has negative surface charges, so it dissolves well in water. Proteins instead have a more complex structure, a carbon skeleton with unevenly distributed charges. DNA solubilizes only in the aqueous phase, and proteins globally are more suited for an apolar solvent. There is, therefore, a phase separation: **DNA** ends in the polar phase, while **proteins** go toward the apolar phase. This separation is done with a single centrifugation. **Fats** go into the apolar phase. RNA should end in the polar phase, but using **basic phenol** (pH>>7) causes hydrolyzation of the RNA (for the extra hydroxyl which has its sugar). DNA is comfortable in a basic environment, being very stable at pH 7-8. Therefore, for DNA extraction, phenol is mixed with a pH 8 buffer; thus, phenol for DNA is obtained. **RNA** should be extracted in water-saturated phenol, at low pH (~4, **acid phenol**). Phenol is heavier than water and therefore tends to go downwards, with a centrifugation it is therefore possible to separate the two phases easily. The solid phenol is liquefied at 60°C and is then stored under the adequate buffer (TrisHCl at pH 8 for DNA, water for RNA). Phenol is usually used together with another apolar solvent, **chloroform**, and with a small quantity of **isoamylic** alcohol as an antifoaming agent.

Following centrifugation, the upper polar, aqueous phase is taken out of the tube. To remove salt excess, and to reduce the sample volume, it is now possible to concentrate DNA by alcohol precipitation.

b) Precipitation. It is used to enrich and to concentrate DNA. Any nucleic acid dissolved in water may be re-precipitated by alcohol. A **salt** sodium (Sodium Acetate) is usually added before extraction or may be added now, to favour the precipitation of DNA molecules, that being negatively charged would repel each other. A cation may neutralize them by binding the phosphate groups.

Ethanol (CH_3-CH_2OH) has an apolar region (CH_3-CH_2) and a polar (OH) region, thus it creates an unstable environment for hydrophilic nucleic acids, favouring their precipitation.

If instead of ethanol **isopropylic alcohol** (isopropanol, 2-propanol) is used half of the dose it is sufficient, because it has three carbon atoms and not two like EtOH: 1 volume of the solution plus 1 volume of isopropyl alcohol. When using ethanol, two volumes of it must be added.

In any case, **cold** alcohol strongly favours precipitation, because low temperature decreases the solubility of the solute. Therefore, it is usual to maintain a bottle of alcohol (ethanol or isopropanol) in the freezer at -20°C, where it will remain liquid.

Thus, following the addition of cold alcohol, and centrifugation in a cold room or a refrigerated centrifuge, a pellet of sodium deoxyribonucleate is collected.

If there is little starting DNA (poor sample), little is found at the end of the extraction: to recover as much as possible nucleic acid, an exogenous source of nucleic acid may be

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

used as a "carrier" to favour precipitation. Yeast tRNA (ala-tRNA) is used, which serves as an aggregation centre and does not affect subsequent analyses. It is estimated that even a few pg of human DNA can be recovered in this way.
c) Washing. It is necessary because precipitated DNA is still very rich in sodium which can cause interference with restriction enzymes or with polymerases. To wash DNA, 70-75% ethanol is used, this solution is very alcoholic but still a little aqueous, so it is still sufficiently apolar to pull down the DNA, but sufficiently polar to retain sodium, "extracting" it from the DNA.
d) Suspension. Following removal of the supernatant, the final DNA pellet is resuspended in one drop (50 µL) of sterile, bidistilled water. In the case of RNA, an RNAse inhibitor may be added, such as RNAsin (extracted from the placenta). Diethylpirocarbonate (DEPC)-treated water was also widely used in the past (however, this substance is toxic).

Note - Due to the presence of an additional hydroxyl group in the sugar (ribose) compared to DNA, RNA is, in general, much more reactive than DNA. Also, it can be cleaved by cellular **RNases** released from cells upon lysis. These enzymes are found on the skin too so that gloves are required during RNA extraction. RNases have strong intrachain disulfide bonds, therefore can be resistant to prolonged boiling and can refold quickly following denaturation. In order to prevent RNA degradation, RNases should be inactivated as rapidly as possible at the very first stage in the extraction process, which should be performed maintaining the tubes on ice and using RNase inhibitors. While RNases do not require divalent cations for their activity, **DNases** do, so the latter can be inhibited by adding ethylenediaminetetraacetic acid (EDTA), a chelating agent, i.e., able to sequester metal ions such as Ca2+. Several types of RNA (e.g., almost all the mRNAs) have a poly(A) tail, that can be used for their selective isolation following binding with poly(T).

2) GLASS BEADS AFFINITY

In **silicate gel** (microscopic glass beads), DNA binds strongly to the beads in the presence of highly concentrated chaotropic agents (it has been known for decades). It is only needed to apply the lysate to these beads packed in a spin-column, and following a centrifugation the DNA remains attached to the beads. With a second washing at different molarity (low osmolarity), the DNA is detached and recovered.
It is a fast method, but less nucleic acid is usually recovered.

3) ION EXCHANGE RESINS

Ion exchange resins (positive) were first available by Qiagen, similar to "miniprep" by Maniatis. These resins exchange an ion with nucleic acids, using the same principle of water purification for domestic use. There are used in the form of spin-columns, to which the sample is applied. This method is also less efficient compared to extraction in organic solvents; however, it allows the experimenter to work on many samples in parallel.

NUCLEIC EXTRACTION CHECKING

The extraction is not finished until it proves that the DNA is there.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Gel electrophoresis is useful to separate DNA molecules because in nucleic acids the charge / mass ratio is constant (**Q/m=k**): each nucleotide brings a net charge.
Best results are obtained when some way to recirculate the buffer between the two chambers of the apparatus (with positive and negative charges) is found.

An **agarose** gel stained with ethidium bromide is prepared.
Ethidium bromide is an intercalating agent selectively staining nucleic acids. It can be visualized under ultraviolet (UV) irradiation, typically at a transilluminator with a 302 nm lamp.
Agarose gel can resolve two bands if they are related to polynucleotides different by at least **12** (in the best case)**-50 bp**. The agarose concentration should be adjusted for the best resolution (higher concentration, up to 2%, resolve smaller molecules, up to 100 bp; low concentrations, down to 0.3%, resolve larger molecules, of thousand bp.

Polyacrylamide gels may resolve nucleic acids molecules varying in size from 10 to 1500 bp and differing even by only **1 nucleotide** (6% total acrylamide), this property is exploited for Sanger sequencing.
Acrylamide monomer is provided as a powder. Unpolymerized acrylamide is a neurotoxin. Upon addition of water, in the absence of oxygen, it polymerizes resulting in the formation of polyacrylamide. Gel size can be regulated by adjusting the concentration of acrylamide.
Bisacrylamide can form cross-links between two acrylamide molecules, thus creating gel pores of regular shape. Polymerization is triggered by a source of free radicals (ammonium persulfate) and a stabilizer (TEMED). The ratio of acrylamide to bisacrylamide is typically 19:1 for the analysis of nucleic acids.
Polyacrylamide gels are usually 1-2 mm thick and are cast vertically between two glass plates, to use a small quantity of this expensive reagent and to remove the air (oxygen), which hampers polymerization. However, systems have been marketed, offering the possibility to pour acrylamide gels horizontally.

DNA gel electrophoresis

1st check - existence: a single band usually contains all genomic DNA, in pieces generated during the extraction by the mechanical forces applied. Simply pipetting, shaking or stirring can shear both strands of DNA, which is chemically inert and durable but physically fragile. Human genomic DNA can thus be otained only in fragmented form, as a collection of fragments originating from random mechanical breakage of the chromosomes, whose mean size depends on the type and the intensity of the forces applied during the extraction process.

2nd check - size: the pieces must be at least 20,000 bases. It can be estimated by comparison with a size marker.

3rd check - quality: DNA must not be degraded, e.g. showing different molecular weight bands scattered throughout the lane.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

4th check - quantity: estimation of the DNA quantity based on the size and brightness of the bands.

The gel is more sensitive than the **spectrophotometer** in doing this estimation, because any bright signal is related to a nucleic acid, while spectrophotometer absorbance at **260 nm** can still detect proteins (although most of them adsorb at **280 nm** wavelength).
A ratio greater than **1.8** between the adsorbance value at 260 nm and the one at 280 nm is also considered as an estimation of a low level of contaminating proteins in the RNA preparation.

5th check - functionality, DNA can be effectively subjected to enzymatic treatment, e.g. digested by a restriction enzyme (e.g. *EcoR1*).

The method recovers any **DNA polynucleotides**, thus, non-human too. If there are viruses or bacteria in the sample, their nucleic acids will also be extracted.

This method can be done on samples of saliva, blood, body fluids; buccal, nasal, pharyngeal, ocular swabs; tissues, hair, sperm.

The final estimate of the quality of the extracted DNA is mainly made on the gel. For the quantitative estimation, it is important to load a "correct" quantity of sample, in fact, the shape of the bands is determined by physical factors, in particular from the quantity of DNA loaded, if this is excessive there is a dragging effect (*trailing*) with the formation of lateral "tails" on the upper edge of the band. If more than 200 ng of DNA are loaded in a 4-4.5 mm wide standard well, these effects begin to be obtained which compromise the definition of the bands, so that less molecular species are distinguished. On the other hand, loading less than 1 ng the bands will be invisible as this is the limit of detection for ethidium bromide. If there is a need to load a lot of DNA or RNA, wider wells should be used (10-tooth comb).

A diploid human cell contains ~7 pg DNA, thus from 1 mL of blood ~40 µg of DNA can be extracted (~6 million white blood cells).
A typical mammalian cell is estimated to contain ~10 pg of RNA. However, the expected recovery of 10 mcg (µg, micrograms) of RNA from 1 million cells may turn into 1/10 of this figure in cell types with a small amount of cytoplasm, because tye main fraction of cellular RNA consists of ribosomal RNA (see below).

RNA gel electrophoresis

Human RNA gel electrophoresis shows three main bands:

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

~5,000 nucleotides: rRNA 28S (5,025 nt)
~2,000 nucleotides: rRNA 18S (1,969 nt)

~100 nucleotides: small rRNAs (5.8S - 159 nt, 5S - 121 nt), tRNAs (<100 nt)

All mRNAs (or noncoding RNAs) are dispersed across the gel lane according to their size, each species being present at low concentration, because total RNA is made up of **85% rRNA**, **10% tRNA**, **1-5% mRNA**.

References

The two most known books about protocols in Molecular Biology are often referred to after the name of their first Authors: Sambrook, and Ausubel, respectively. A shared feature is the discussion of the basic principles of the procedures presented.

Molecular Cloning: A Laboratory Manual
The first one-volume edition (1982) was authored by the molecular biologist Tom Maniatis, it has been known for many years as "The Maniatis" and has been present in virtually any laboratory of Molecular Biology and Genetics.
Subsequent three-volume editions were authored by Sambrook, Fritsch and Maniatis (1989) and by Sambrook and Russell (2001). Currently, the fourth edition is available (**Green and Sambrook, 2012**) at the Cold Spring Harbor Laboratory (CSHL) publisher.

Current Protocols in Molecular Biology (popularly referred to as "The Red Book") by **Ausubel, Brent, Kingston, Moore, Seidman, Smith and Struhl** consists of three looseleaf volumes and has been published by Wiley in 1988. Quarterly updates can be filed into the looseleaf.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

NUCLEIC ACID HYBRIDIZATION PRINCIPLES

"Blotting" means transferring one thing to another in one fell swoop, e.g. to blot some ink onto a paper.
In 1975 Edwin Southern proposed to hybridize nucleic acids immobilized on a solid support. Until then, the hybridization had been performed only in solution (two chains pairing with each other to give a double helix in a test tube). The solid support (nitrocellulose, at the time) allowed maintaining a permanent copy of the molecules "recorded" in the position they had taken in the electrophoretic run, following the transfer of the molecules in the gel to the piece of nitrocellulose.
In order to identify specific DNA fragments, it had to be broken up.
DNA extraction and digestion, gel separation of the fragments and their transfer on a solid support, hybridization with a complementary probe to the specific sequence of interest are the main steps of the method. The stability of the bond between the probe and the target molecules depends on the percentage of similarity.
Hybridization can happen by base-pairing between any type of nucleic acid strand: DNA-DNA, hybrid DNA-RNA, RNA-RNA.

SOUTHERN BLOT FOR DNA

1. DNA enzymatic digestion. If we transfer undigested DNA, we would have molecules all of the same length, and therefore a single band would form. It is useful in this case to use restriction endonucleases able to cut at the inside of a DNA molecule. DNase is an aspecific endonuclease. "Restriction" means that the enzyme cuts at specific points. They have been detected in bacteria. After extraction, we have DNA fragments about 20-50 kb long. The restriction enzyme must have a cut-off frequency lower than the size of the molecules. So-called frequent cutters are used, they are enzymes that make frequent cuts on average in the genome. *Eco*RI recognizes a sequence of 6 bases.
We need:
- Purified and concentrated DNA substrate. If not purified, it may contain restriction enzyme inhibitors.
- An active enzyme.
- Optimal reaction conditions.
- Buffer with optimal saline concentration.

One Unit of the restriction enzyme is meant to be the amount of enzyme that digests 1 µg of DNA in one hour.
An **over-digestion** is usually desired (by increasing the reaction time or the amount of enzyme present) which ensures that all the sites are cut.
Class II enzymes are used (those recognizing shorter 4 base sequences cut too much). Restriction mixtures are generally made to have an optimal cut-off frequency. *Not*1 is a rare-cutter, recognizing a sequence of 8 bp.
Southern blot tells us if there is a gene, and how many copies there are compared to normal control.
Enzymes must be diluted very carefully because protein molecules in water tend to

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

undergo denaturation. All enzymes are mixed with BSA (Bovine Serum Albumin), which is a stabilizer, and it does not allow the solution to be too diluted. They are stored at -20°C with 50% glycerol. Freezing damages the protein structure, because ice crystals break the protein molecule. Glycerol has a lower freezing point than water. At -20°C, the enzyme is more stable, while glycerol does not freeze.
The glycerol, in the final solution, must be at most 5%, an excess of it can give rise to phenomena such as "star activity": the enzyme can cut sequences that are not its specific target. If the digestion time increases, the amount of enzyme can be reduced. The enzyme is a catalyst, it is not lost or consumed. The functional efficacy of the enzyme can be checked during the work with a minigel.
Once the reaction has taken place, a standard agarose gel is made (1% in general) and a streak of bands is obtained.

The probe binds to all the complementary pieces of DNA that are on the gel, so more bands will be highlighted. When we have the truly complete DNA sequence, with "restriction simulators" (they are based on word processors: they can find sequence "text") it is possible to predict how a given gene will be cut by a particular mixture of restriction enzymes, except for polymorphisms.
There are restriction enzymes that cut the two DNA strands symmetrically (generating "blunt" DNA ends) or asymmetrically (generating "sticky" ends, with one strand protruding over the other).

2. Agarose gel electrophoresis of the DNA fragments.

3. Transfer (blotting) of the separated DNA fragments from the gel **to a solid support** (a sheet filter of **nitrocellulose**, or better of **nylon**, which is more robust and may be electrically (positively) charged to better retain DNA.).
a) The classic blot is a **capillary blot**: sheets of paper soaked in SSC (sodium saline citrate, a very concentrated solution of NaCl and Na citrate) attract the gel water that drags up the molecules that remain trapped in the filter that has molecular (sized) pores. The molecules rise in a straight line and get trapped in the same position they occupy in the gel.
If you don't stack everything well, it may fail.
b) Reverse blot: the gel is placed on the membrane, in this way, gravity is also exploited, and the saline solution is continuously dispensed from a wet sponge at the top of the blot stack.

In order to keep the DNA molecules denatured, so that they can be able to hybridize with the probe, the blotted filter can be soaked in an alkaline solution, or blot can be directly performed using NaOH 0.4 M as the transfer medium.

4. Fixation of the DNA to the filter. By using a **UV** lamp at 254 nm, at a certain distance from the filter and for 5 minutes, the chemical groups of sugar and PO_4 are energized so that they bind to the amino groups -NH2 of the nylon membrane. Our standard exposure is 12 watts, about 15 cm, for 5 minutes. Following fixation, the

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

blotted membrane can be hybridized up to 12 times in our hands. If it is fixed too much, however, due to the formation of too many bonds with the filter, the bases are made unavailable for hybridization. The same is also obtained by providing temperature: putting the filter in the microwave for a few minutes.
A nylon sheet is obtained with all the lanes and here the fragments of DNA separated by size.

5. Probe labelling

A) Random priming

The easiest way is to use an in vitro polymerase that copies the filaments using a radioactive nucleotide: ^{32}P dCTP.
In the test tube: cDNA probe, random primers, in order to prime synthesis of the labelled strand from any template sequence, **DNA polymerase**, the 4 dNTPs, one of which labelled in one of the three phosphate groups, which are called according to their proximity to the sugar from alpha to gamma:

```
base-sugar-P-P-P
           α  β  γ
```

An α-labelled ^{32}P is purchased because, in the polymerization reaction, only alpha P remains in the synthesized strand. It must be kept in mind that "alpha" does not indicate here the type of emission of ^{32}P, which is a β-emitter (fast electrons), but the position of the radioactive P.

B) Oligo terminal (3´) labelling

If a ready-made single-strand oligonucleotide is the probe to be labelled, it will not have the 5´phosphate that natural molecules have, instead, it has a classic **3´-hydroxyl end**.
The **terminal deoxynucleotidyl transferase (TdT)** is used in this case (recombinant TDT). This enzyme transfers deoxynucleotides to the terminal position at the extreme 3´. It is an example of a polymerase (an enzyme that extends a DNA polymer by catalyzing the phosphodiester bond) that does not need a template. It is found only in pre-lymphocytes (a marker for the diagnosis of leukaemia) and works in the presence of divalent cations: if the added nucleotide is a purine, it is better to use Mg^{++}, if it is a pyrimidine it is better to use Co^{++}. Single nucleotides can be added using deoxynucleotides. In this case, in the reaction tube we put:
- the oligonucleotide to be marked,
- **dATP** where the alpha P is a beta-emitting ^{32}P (ATP because it seems that the TdT has a greater affinity for the A),
- the TdT.

A chain of radioactive As will be added at the 3 ´ end of the oligo molecules.

C) Oligo 5´ labelling

By using a gamma ^{32}P deoxynucleotide and the enzyme **PNK (Polynucleotide Kinase)**, the gamma phosphate may be transferred to the **5´ end** of a single strand probe whose

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

original 5′ phosphate group has been removed by **alkaline phosphatase**. In this way, a single radioactive atom is added to each probe molecule, so that this is a weaker labelling.

6. Hybridization. Hybridization consists in making the filter react with a probe to verify the existence, quantity and possible alterations of a particular fragment of DNA.
Steps:
- **Probe** preparation (in solution). The probe must be heated at 94-100°C to **denature** it before adding it to the hybridization reaction.
- **Renaturation** kinetics: the double helix is reformed between the probe and the target. There is the mixing of two single strands that make a double helix that was not there before. It is an indirect mode, the probe must be known and must also be visible (for us or for the detection instrument) in order to reveal an unknown and invisible molecule.
In the presence of adequate **salt** concentration (cations, to neutralize negatively charged DNA molecules allowing their pairing) and **temperature** conditions, only specific binding of the probe to the target will be favoured.
Under "stringent" hybridization conditions, one could also distinguish between target sequences divergent only for one base.
If "permissive" conditions are chosen, similar genes will be identified (members of gene families), slightly different from the known probe, also across different species. DNA and RNA have extraordinary flexibility to make single helix sections alternating with double helix sections.
Thus, any molecule that has a certain homology with the probe can be recognized, even if for example there are some blocked bases of the target.
Cocktails of oligonucleotides may be used to hybridize with different exons of the same gene.

7. Washes. Removal of all that is weakly bound. In the classic case, we use a probe as specific as possible.
Wash stringency may be adjusted by rising **temperature** and/or
lowering **salt** concentration: both these conditions tend to detach any nonspecifically bound probe molecules from the hybridized filter.
As an extreme case, when all the probe must be removed in order to start a new hybridization experiment with a different, the blot membrane is boiled in distilled water (the highest temperature, in absence of salts).

8. Autoradiography. It highlights the base pairing between the probe and the target. A photosensitive film covered with a photographic (radiographic) emulsion made of silver halide is used. The **beta particles** emitted by the labelled probe hit the electrons of this substance and cause the **silver** to release and form a precipitate, a black dot that we can see on the developed film. ^{32}P emits beta particles with high energy content. The hybridized, radioactive filter is wrapped in a plastic sheet to avoid contamination. The silver precipitate is directly proportional to the number of beta particles emitted;

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

however, there is a slight widening of the bands, a blur around the point of origin, this phenomenon can be a problem for nearby bands that can overlap. To improve autoradiography, we can:
1) add an **intensifying screen**. It is screen coated with calcium tungstate. The most energetic beta particles that manage to cross the film hit the screen which covers the inside of the cassette, this generate new particles that come back and hit the film again. Sensitivity increases but the blur increases.
2) place the autoradiographic cassette in the **freezer** at -20°C (or even in the freezer -at 80°C). The reaction that occurs in a radiography is the precipitation of silver, and any precipitation is favoured at low temperatures.

NORTHERN BLOT FOR RNA

It is useful for studying the expression of a gene: in what tissues and periods of life is it active, how active is it, etc.
The general procedure in similar to Southern Blot, with some variation that will be highlighted.

Ribosomal RNA (rRNA) may be a useful reference point.
If there is degradation, the 28S tends to break into two pieces that end in the 18S-band, in this case, the brilliance of the 18S-band increases and that of the 28S decreases.
When working with rRNA to obtain a good resolution, a maxigel is made with larger wells to be able to load more RNA in order to have an appreciable signal (10 to 20 µg).
The average quantity of RNA per cell is about 10 pg; therefore, it is necessary to start from about 1 x 10^6 cells. From 10 mL of blood, about 10 µg of RNA are obtained; from other tissues, it is easier to have larger quantities.

RNA is run on the gel as much as possible to separate the bands well. **Formaldehyde** is added in the gel to inhibit the formation of hydrogen bonds. If it is not used, single-stranded filaments tend to fold due to self-complementarity. RNA has a very strong tendency to make double helices in complementary regions.
The Northern gel is blotted (in neutral citrated saline solution because the denaturant is already there: it is formaldehyde) on a nylon membrane. Ethidium bromide is also loaded into the well to be able to highlight the **28S** and **18S** bands for UV brilliance. A mark can be made by a pencil in correspondence the centre of the two bands, thus obtaining two points corresponding to molecules of known length and allowing building a straight line on which to evaluate the other bands.

There is a second possibility which consists in separating the polyadenylated messengers after removing the ribosomal RNA. On the filter only the putative messengers remain, the problem is that if there is also a trace of ribosomal RNA, there is the risk of having an unspecific signal for a cross-hybridization (cross-hybridization), even a low homology is enough because many copies of ribosomal not eliminated give a signal (bright band). For marker, a known marker is used. The rRNA release procedure also reduces the number of molecules that interest us: at least 50-100 µg of starting is

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

required (e.g. 4 flasks of cells). We can hybridize the membrane as for Southern and then do autoradiography. We can expect to see even completely white wool if the cells do not express that messenger, for this reason, some control points are needed: positive controls: at least one known RNA (e.g. 2 Kb) certainly present in the sample, indicates that the method has worked.

INTERPRETATION OF NORTHERN DATA

Reference markers are used: **"housekeeping" genes** (the housekeepers, who manage the house) are not specific to certain cells, they are active in all cells, and have two fundamental characteristics for their use in this sense:
1) are always expressed;
2) have a constant, and high, level of expression.
Northern manages to see RNA up to 5-10 copies per cell. The housekeeping gene must be expressed far more than 5-10 copies per cell.
The most used genes are:
- BETA-ACTIN (*ACTB*) mRNA (cytoskeleton).
- β2 MICROGLOBULIN (*B2M*) is expressed in all nucleated cells.
- GLUCOSE-3-PHOSPHATE-DEHYDROGENASE (G3PDH) is an enzyme involved in basic carbohydrate metabolism.

Housekeeping genes can be used in a Northern quantity to calibrate the expression of the gene on the quantity, for example of the actin band. These genes have also been found to have variations in expression as a function of the cell cycle phase. *B2M* appears to be less affected by the cell cycle. However, it is less expressed.
By using Northern blot, kinetic studies can be done: at what time after stimulation is the peak of mRNA.

Additional bands occurring besides the mRNA expected band could be due to:

1. Primary, immature transcripts.
They are generally processed quickly, however sometimes it can be seen. It has many regions that are not there in the mature (intronic regions), and it can be seen in the intermediate stage of processing (the processing kinetics is very different).
2. Alternative splicing isoforms,
also hybridizing to the probe.
3. Member of the same gene family,
whose mRNAs partly hybridize with a probe recognizing a similar member of the same family.
4. Cross-hybridization
with 28S or 18S RNA, seen in all all lanes if there is partial homology with these molecules very abundant in total RNA (85%).

An example of a Northern blot to characterize a newly discovered gene is here.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

If the size of the expected mRNA is known a priori, and these size are very different, a filter can also be hybridized with several probes simultaneously.
Some vendors sell ready-made blots with 12 different human tissues, ready to be hybridized with a probe.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari
www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari
www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

© 1999 Nature America Inc. • http://genetics.nature.com

Expression profiling using cDNA microarrays

David J. Duggan, Michael Bittner, Yidong Chen, Paul Meltzer & Jeffrey M. Trent

Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. e-mail: jtrent@nhgri.nih.gov

cDNA microarrays are capable of profiling gene expression patterns of tens of thousands of genes in a single experiment. DNA targets, in the form of 3′ expressed sequence tags (ESTs), are arrayed onto glass slides (or membranes) and probed with fluorescent- or radioactively-labelled cDNAs. Here, we review technical aspects of cDNA microarrays, including the general principles, fabrication of the arrays, target labelling, image analysis and data extraction, management and mining.

© 1999 Nature America Inc. • http://genetics.nature.com

Ambitious projects aimed at cloning, mapping and sequencing the genomes of various organisms, including that of *Homo sapiens*, have been launched worldwide. In all cases, the fruits of these labours will provide a solid platform from which to attempt the larger goal of understanding how genomes result in the organisms they specify. The success of these international efforts is impressive. So far, complete genomic sequences of 17 organisms, including the eukaryote *Saccharomyces cerevisiae*, have been produced. The mapping (both genetic and physical) and sequencing phases of the Human Genome Project are ahead of schedule. Researchers have catalogued more than 1.1 million expressed sequence tagged sites (ESTs), corresponding with 52,907 unique human genes[1] (www.ncbi.nlm.nih.gov/UniGene). However, the function, expression and regulation of more than 80% of them has yet to be fathomed. The next phase of the human genome project will place strong emphasis on assigning function to these genes.

The ability to identify genes at the nucleic acid level rather than proceeding from a known protein to its chromosomal counterpart has prompted efforts to likewise extract functional information at the nucleic acid level. Two methods are currently in use. The 'sequence' approach has led to the discovery of a wide variety of sequence motifs encoding structural domains, such as DNA-binding and nucleotide-binding domains[2], thus providing clues to gene function. Another route for exploring the function of a gene is by determining its pattern of expression. The accumulation of expression data has yet to reach the point at which it is possible to speak of expression motifs, but it does suggest that this is a plausible outcome of the approach[3–5].

Various methods are available for detecting and quantitating gene expression levels, including northern blots[6], S1 nuclease protection[7], differential display[8], sequencing of cDNA libraries[9,10] and serial analysis of gene expression[11] (SAGE). Augmenting this coterie are two array-based technologies—cDNA and oligonucleotide arrays. These allow one to study expression levels in parallel[3,12,13], thus providing static information about gene expression (that is, in which tissue(s) the gene is expressed) and dynamic information (that is, how the expression pattern of one gene relates to those of others). The high degree of digital data extraction and processing of these techniques supports a variety of samples or experimental conditions.

Although both cDNA and oligonucleotide arrays are capable of analysing patterns of gene expression, fundamental differences exist between the methods. Here, we focus primarily on technical aspects of cDNA microarrays, although some comparison with the oligonucleotide array (see page 20 of this issue (ref. 14)) will be made where appropriate.

Principle of method

As reviewed by Ed Southern on page 5 of this issue, hybridization between nucleic acids (one of which is immobilized on a matrix) provides a core capability of molecular biology[15]. This method provides high sensitivity and specificity of detection as a consequence of exquisite, mutual selectivity between complementary strands of nucleic acids. Historically, most applications of this method have employed a single, pure, labelled oligonucleotide or polynucleotide species in the liquid phase and complex mixtures of polynucleotides attached to a solid support. Transcript abundance is assayed by immobilizing mRNA or total RNA (electrophoretically separated or in bulk) on membranes and then incubating with a radioactively labelled, gene-specific target. If multiple RNA samples are immobilized on the same matrix, one obtains information about the quantity of a particular message present in each RNA pool.

cDNA arrays alter this strategy in several ways (Fig. 1). In an array experiment, many gene-specific polynucleotides derived from the 3′ end of RNA transcripts are individually arrayed on a single matrix. This matrix is then simultaneously probed with fluorescently tagged cDNA representations of total RNA pools from test and reference cells, allowing one to determine the relative amount of transcript present in the pool by the type of fluorescent signal generated. Relative message abundance is inherently based on a direct comparison between a 'test' cell state and a 'reference' cell state; an internal control is thus provided for each measurement (Fig. 2). The scheme is similar when using radiolabelled probe, but it is not possible to carry out simultaneous hybridization of test and reference samples. In such cases, serial or parallel hybridization is required, introducing the possibility of higher variability in comparisons of expression level.

The adaptable nature of the fabrication and hybridization methods allows the technique to be applied widely—the only limitations are the availability of clones for the solid phase and the quality of RNA samples derived from the cells (or tissues) to be compared. This is illustrated by diverse applications that include: investigating gene expression in the roots and leaves of *Arabidopsis thaliana*[3], human T cells exposed to phorbol ester[12], rheumatoid arthritis and inflammatory bowel disease[16], tumorigenic versus non-tumorigenic cell lines[4], the diauxic shift from anaerobic to aerobic metabolism in *S. cerevisiae*[5,17] (yeast),

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

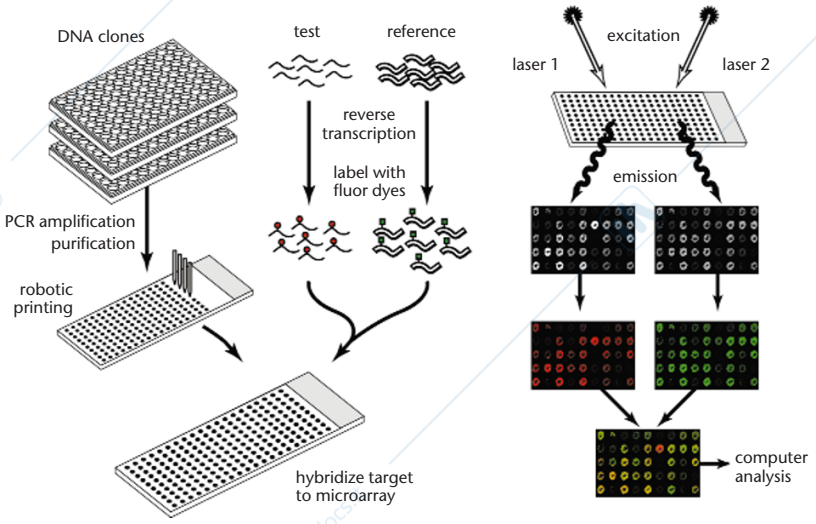


Fig. 1 cDNA microarray schema. Templates for genes of interest are obtained and amplified by PCR. Following purification and quality control, aliquots (~5 nl) are printed on coated glass microscope slides using a computer-controlled, high-speed robot. Total RNA from both the test and reference sample is fluorescently labelled with either Cye3- or Cye5-dUTP using a single round of reverse transcription. The fluorescent targets are pooled and allowed to hybridize under stringent conditions to the clones on the array. Laser excitation of the incorporated targets yields an emission with a characteristic spectra, which is measured using a scanning confocal laser microscope. Monochrome images from the scanner are imported into software in which the images are pseudo-coloured and merged. Information about the clones, including gene name, clone identifier, intensity values, intensity ratios, normalization constant and confidence intervals, is attached to each target. Data from a single hybridization experiment is viewed as a normalized ratio (that is, Cye3/Cye5) in which significant deviations from 1 (no change) are indicative of increased (>1) or decreased (<1) levels of gene expression relative to the reference sample. In addition, data from multiple experiments can be examined using any number of data mining tools.

murine T cells challenged with 4-phorbol-12-myristate-13-acetate[13] and in *Streptococcus pneumoniae*[18].

Fabrication

Production of arrays begins with the selection of the 'probes' to be printed on the array. In many cases, these are chosen directly from databases including GenBank (ref. 19), dbEST (ref. 20) and UniGene (ref. 1), the resource backbones of the array technologies (see page 25 of this issue (ref. 21)). Additionally, full-length cDNAs, collections of partially sequenced cDNAs (or ESTs), or randomly chosen cDNAs from any library of interest can be used. Arrays for higher eukaryotes are typically based on the EST portions of these projects, whereas for yeast and prokaryotes, probes are usually generated by amplifying genomic DNA with gene-specific primers. Given the expense of obtaining clones, producing DNA from them, and printing them, it is usually preferable to produce arrays with a low redundancy of representation, so as to survey the broadest possible set of genes.

In this regard, the human UniGene database represents an excellent model of the kind of informational base one needs both to choose clones and to evaluate expression profiles. It includes a summary of information about the function of a particular gene, its genomic location, clones that contain the gene and connections to other relevant databases and literature sources. On the other hand, no other organisms have such a well-developed EST database, a limitation, given that cDNA microarrays also permit the 'assay' of uncharacterized cDNAs (which may represent genes with informative expression patterns).

cDNA arrays are produced by spotting PCR products (of approximately 0.6–2.4 kb) representing specific genes onto a matrix. These are usually generated from purified templates, so that cellular contaminants do not find their way onto the array. Typically, the PCR product is partially purified by precipitation, gel-filtration, or both —to remove unwanted salts, detergents, PCR primers and proteins present in the PCR cocktail. For both glass and membrane matrices, each array element is generated by the deposition of a few nanoliters of purified PCR product, typically of 100–500 µg/ml (see page 18 of this issue (ref. 22)) Printing is carried out by a robot that spots a sample of each gene product onto a number of matrices in a serial operation. The first spotting robots relied on contact printing with a device not unlike a fountain pen. Many variations on this design are now available (see page 31 of this issue (ref. 21)), in addition to a 'spotter' that is essentially a capillary tube, to which a low but constant pressure is applied. Non-contact printing modes, using either piezo or ink-jet devices, are also being evaluated.

The types of membranes commonly used are nitrocellulose and charged nylon commercial varieties that are used for various blotting assays. Glass-based arrays are most often made on microscope slides, which have low inherent fluorescence. These are coated with poly-lysine, amino silanes or amino-reactive silanes[12], which enhance both the hydrophobicity of the slide and the adherence of the deposited DNA. They also limit the spread of the spotted DNA droplet on the slide.

In most cases, DNA is cross-linked to the matrix by ultraviolet irradiation. After fixation, residual amines on the slide surface are reacted with succinic anhydride to reduce the positive charge at the surface. As a final step, some percentage of the DNA deposited is rendered single-stranded by heat or alkali (see page 19 of this issue for a detailed description of procedures[22]). The state of bound DNA is ill-defined. It is deposited in double-stranded form, intra-strand cross-linked to some extent, and may well have multiple constraining contacts with the matrix along its length (induced by drying the DNA onto the matrix; Fig. 3). It is therefore probably not the best hybridization probe. One can imagine that oligonucleotide matrices, with their short chains and single points of constraint at each chain end, may well be a far more accessible probe for hybridization. Against this advantage, however, must be weighed the disadvantages of using short-chain detectors. Chief among these are the variations in melting temperature due to AT–GC composition, and the reduction in specificity due to truncating the number of nucleotides from hundreds to as few as twenty. A format in which the accessibility of a simply tethered, single-stranded probe could be combined with the specificity of a long probe would provide a considerable improvement for the field.

Target labelling and hybridization

The targets for arrays are labelled representations of cellular mRNA pools. Typically, reverse transcription from an oligo-dT primer is used. This has the virtue of producing a labelled product from the 3′ end of the gene, directly complementary to immobilized targets synthesized from ESTs. Frequently, total RNA pools (rather than mRNA selected on oligo-dT) are labelled, to maximize the amount of message that can be obtained from a given

© 1999 Nature America Inc. • http://genetics.nature.com

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

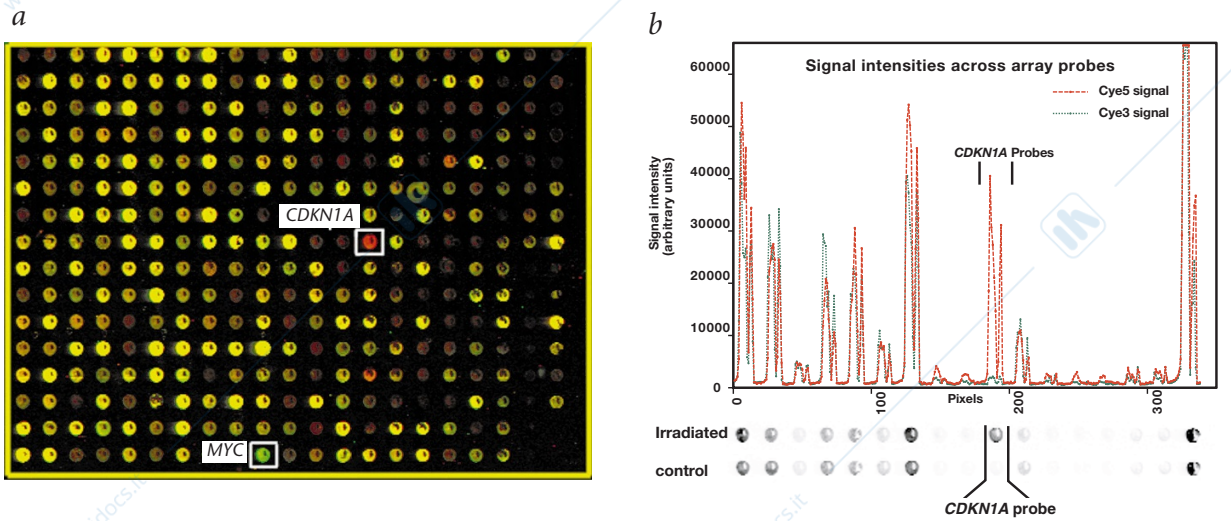


Fig. 2 Quantitations from two-colour hybridization. ***a***, A segment of an array to which targets from γ irradiated ML1 cells (red) and untreated ML1 cells (green) are hybridized. Highly differential hybridization is visible at the detectors for *CDKN1A* and *MYC* (boxed). ***b***, Intensity along a horizontal axis running through *CDKN1A* and several detectors on either side. The intensity profiles are nearly coincident at each probe except *CDKN1A*. At *CDKN1A*, the signal from the uninduced cells is near the threshold of detection, whereas the signal from the induced cells is considerably greater.

amount of tissue. The purity of RNA is a critical factor in hybridization performance, particularly when using fluorescence, as cellular protein, lipid and carbohydrate can mediate significant non-specific binding of fluorescently labelled cDNAs to slide surfaces. For radioactive detection, ^{33}P dCTP is preferred to more energetic emitters, as array elements are physically close to each other and strong hybridization with a radioactive target can easily interfere with detection of weak hybridization in surrounding targets. As fluorescent labels, Cye3-dUTP and Cye5-dUTP are frequently paired, as they have relatively high incorporation efficiencies with reverse transcriptase, good photostability and yield, and are widely separated in their excitation and emission spectra, allowing highly discriminating optical filtration.

A clear limitation to the application of this technology is the large amount of RNA required per hybridization. For adequate fluorescence, the total RNA required per target, per array, is 50–200 μg (2–5 μg are required when using poly(A) mRNA). For mRNA present as a single transcript per cell (approximately 1 transcript per 100,000), application of target derived from 100 μg of total RNA over an 800 mm^2 hybridization area containing 200-μm diameter probes will result in approximately 300 transcripts being sufficiently close to the target to have a chance to hybridize. Thus, if the fluorescently tagged transcripts are, on average, 600 bp, have an average of 2 fluor tags per 100 bp and hybridize—all of them—to their probe, approximately 12 fluors will be present in a 100-μm^2 scanned pixel from that probe. Such low levels of signal are at the lower limit of fluorescence detection, and could easily be rendered undetectable by assay noise. Although radioactive targets may have a higher intrinsic detectability, they too reach a level of dilution that prohibits effective detection, thus precluding experimentation on very small numbers of cells (Fig. 4).

A variety of means by which to improve signal from limited RNA has been proposed. These are being evaluated by our laboratory and many others. Efficient mixing of the hybridization fluid should bring more molecules into contact with their cognate probe, increasing the number of productive events. This entails, however, a larger 'mixing' volume, which might offset the potential gain. Methods that produce multiple copies of mRNA using highly efficient phage RNA polymerases have been developed[23]. A version of this approach, in which labelled target (cRNA) is made directly from a cDNA pool, having a T7 RNA polymerase promoter site at one end via *in vitro* transcription, has been applied to arrays[13]. Post-hybridization amplification methods have also been reported in which detectable molecules are precipitated at the target by the action of enzymes 'sandwiched' to the cDNA target[24]. Detection of hybridized species using mass spectroscopy or local changes in electronic properties can also be imagined[25,26].

Image analysis and data extraction

The highly regular arrangement of detector elements and crisply delineated signals that result from robotic printing and confocal imaging of fluor-detected arrays renders image data amenable to extraction by highly developed, digital image processing procedures. Grids specifying target locations can be readily overlaid on the images. Local sampling of background can be used to specify a threshold which true signal must exceed. Mathematical morphology methods can be used to predict the likely shape and

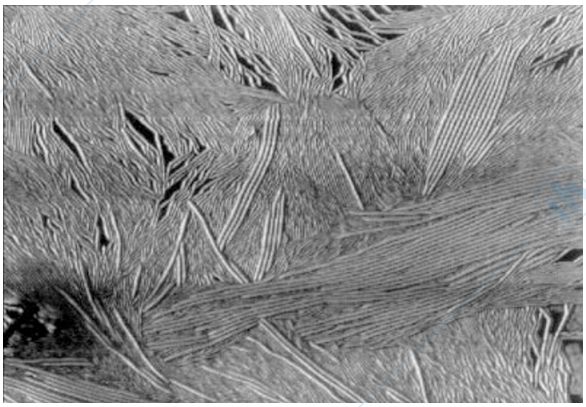


Fig. 3 Atomic force microscopy of DNA on a microarray. This is a micrograph of a portion of a hybridization probe from a yeast microarray, taken after the array was subjected to hybridization. The DNA is clearly deposited at a sufficient density to allow many kinds of strand-to-strand interactions. The width of the picture represents a scanned distance of 2 μm. Image kindly provided by J. DeRisi (Stanford) and E. Carr (Hewlett-Packard).

© 1999 Nature America Inc. • http://genetics.nature.com

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

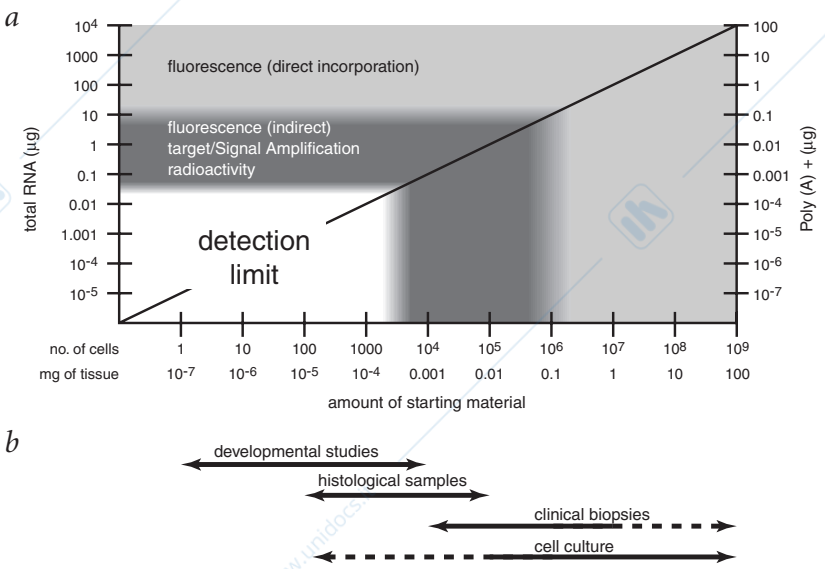


Fig. 4 Detection schemes and applications of cDNA microarrays. Quantitative changes in gene expression can be detected using several schemes for which the limits of detection vary (***a***). Direct incorporation of fluorescent nucleotides into the cDNA target can be used to examine expression profiles from 10 µg or more of total RNA. Indirect fluorescence, as well as target and signal amplification and radioactivity, on the other hand, can be used to detect expression profiles from as little as 50 ng of total RNA. This detection limit allows for the investigation of expression profiles from numerous biological sources including cell culture, clinical biopsies (including autopsy material) and histological samples (***b***). Improvements in technology will permit the detection of expression profiles from less than 50 ng of total RNA, increasing the utility of the technology with respect to studies in development. The limits of the various techniques are constantly changing, and this chart is meant only to illustrate of current performance levels.

placement of the hybridization signal. By applying these methods it is possible to accurately detect even weak signals[27] and extract a mean intensity above background for the target. In contrast, extraction of data from film or phosphor-image representations of radioactive hybridizations presents many difficulties for image analysis. If the array is on a membrane, there is frequently nonlinear warping of the matrix, which means that the observed array will not have the strict geometric regularity of an array printed to a stiff matrix, such as glass. This introduces difficulty in developing highly accurate grids to specify target locations. The spread of detectable particles from a disintegrating nuclide to the detector is highly sensitive to variations in distance between source and detector, and produces a smooth transition from the highest levels of intensity to background. This ensures that the image produced by radioactive exposure is composed of sections at many focal planes, and renders impossible the application of single, simple, point-spread functions to reconstitute a 'focused' representation of the data. The smoothness of the transition from maximum signal intensity to background signal intensity makes consideration of local background for each signal a difficult proposition as one does not observe an abrupt, readily discerned transition between signal and background, but a smooth curve without a sharp derivative.

In carrying out comparisons of expression data using measurements from a single array or multiple arrays, the question of normalizing data arises. All experiments are carried out under conditions of a large excess of immobilized probe relative to labelled target. The kinetics of hybridization are therefore pseudo-first order, and inter-probe competition is not a factor. Under these conditions, the linear differences arising from exact amount of applied target, extent of target labelling, efficiencies of fluor excitation and emission, and detector efficiency can be compounded into a single variable and the information from each detection channel normalized. It is best to achieve normalization by adjusting the sensitivity of detection (photomultiplier voltage with fluorescence or exposure time with radioactivity) so that the measurements occupy the same dynamic range in the detector. There are essentially two strategies that can be followed in carrying out the normalization. One is based on a consideration of all of the genes in the sample, and the other, on a designated subset expected to be unchanging over most circumstances. In either case, variance of the normalizing set can be used to generate estimates of expected variance, leading to predicted confidence intervals. In instances of closely related samples, the transcript level of many genes will remain unchanged, making global normalization a useful tool. As samples become more divergent, the fraction of genes showing altered transcript levels increases, and global normalization yields a poorer estimate of normalization than would be achieved using a subset of constantly expressed genes. Explicit methods have been developed which make use of a subset of genes for normalization, and extract from the variance of this subset statistics for evaluating the significance of observed changes in the complete dataset[27].

An aspect common to all array techniques is the extent of reliability and variance in measurements. So far, most array methods have been validated by probing northern blots of the biological samples. As with sequencing, the best comparisons and measures of reliability can be made only when large data sets containing significant repetitions and overlapping data are freely available. One can, however, clearly envisage strengths and weaknesses. The simple and highly determined nature of immobilized hybridization probes in oligonucleotide arrays make them likely to yield the highest level of reproducibility of absolute measurement for a given element. The ability of cDNA arrays to achieve element-by-element normalization with two-colour fluorescence detection and to use a single, highly specific immobilized probe could provide the most accurate measurements of relative expression levels. All methods should readily disclose large changes in transcript levels among those genes readily detected.

Data management and mining

All array methods require the construction of databases for the management of information on the genes represented on the array, the primary results of hybridization and the construction of algorithms to make it possible to examine the outputs from single and multiple array experiments (ref. 27; see also, page 51 of this issue (ref. 28)). Methods applied to microarray data analysis have essentially been correlation-based approaches that apply methods developed for the analysis of data which are more highly constrained (such a protein or amino acid sequence comparisons) than at the transcript level. This level of analysis on large data sets could provide new perspectives of the operation of genetic networks. Comparison of expression profiles will undoubtedly pro-

© 1999 Nature America Inc. • http://genetics.nature.com

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

vide useful insights into the molecular pathogenesis of a variety of diseases (ref. 29; see also, page 48 of this issue (ref. 30)). It will not, however, deliver the kind of intimate understanding of the highly inter-related control circuitry that is necessary to achieve true understanding of genome function. A number of recent publications suggest that to achieve this objective, we should reconsider our perception of transcriptional control as a simple on-off switch to a model whereby control is analogous to a highly gated logic circuit, where numerous, often contradictory, inputs are summed to produce a response[31–33]. To reach these goals, biologists must expand the arsenal of tools they use to analyse expression data—recruiting statisticians and mathematicians to consider multivariant problems of a size never before attempted.

Acknowledgements

A host of talented investigators have contributed to the NIH Microarray Project, including: Y. Jiang, A. Glatfelter, G. Gooden, J. Kahn, M. Boguski, G. Schuler, O. Ermolaeva, E. Dougherty, T. Pohida, P. Smith, S. Leighton, J. Hudson, A. Fornace, S. Amundson, S. Zeichner, C. Xiang, R. Simon, J. DeRisi & P. Brown.

1. Schuler, G.D. *et al.* A gene map of the human genome. *Science* **274**, 540–546 (1996).
2. Henikoff, S. *et al.* Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**, 609–614 (1997).
3. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
4. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* **14**, 457–460 (1996).
5. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
6. Alwine, J.C., Kemp, D.J. & Stark, G.R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl Acad. Sci. USA* **74**, 5350–5354 (1977).
7. Berk, A.J. & Sharp, P.A. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell* **12**, 721–732 (1977).
8. Liang, P. & Pardee, A.B. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 967–971 (1992).
9. Adams, M.D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).
10. Okubo, K. *et al.* Large-scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* **2**, 173–179 (1992).
11. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
12. Schena, M. *et al.* Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA* **93**, 10614–10619 (1996).
13. Lockhart, D.J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).
14. Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nature Genet.* **21**, 20–24 (1999).
15. Southern, E. Mir, K. & Shchepinov, M. Molecular interactions on microarrays. *Nature Genet.* **21**, 5–9 (1999).
16. Heller, R.A. *et al.* Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl Acad. Sci. USA* **94**, 2150–2155 (1997).
17. Wodicka, L., Dong, H., Mittmann, M., Ho, M.H. & Lockhart, D.J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.* **15**, 1359–1367 (1997).
18. de Saizieu, A. *et al.* Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. *Nature Biotechnol.* **16**, 45–48 (1998).
19. Benson, D.A., Boguski, M.S., Lipman, D.J. & Ostell, J. GenBank. *Nucleic Acids Res* **25**, 1–6 (1997).
20. Boguski, M.S., Lowe, T.M. & Tolstoshev, C.M. dbEST—database for "expressed sequence tags". *Nature Genet.* **4**, 332–333 (1993).
21. Bowtell, D.L. Options available—from start to finish—for obtaining expression data by microarray. *Nature Genet.* **21**, 25–32 (1999).
22. Cheung, V.G. *et al.* Making and reading microarrays. *Nature Genet.* **21**, 15–19 (1999).
23. Phillips, J. & Eberwine, J.H. Antisense RNA amplification: a linear amplification method for analyzing the mRNA population from single living cells. *Methods* **10**, 283–288 (1996).
24. Chen, J.J. *et al.* Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics* **51**, 313–324 (1998).
25. Thorp, H.H. Cutting out the middleman: DNA biosensors based on electrochemical oxidation. *Trends Biotechnol.* **16**, 117–121 (1998).
26. Marshall, A. & Hodgson, J. DNA chips: an array of possibilities. *Nature Biotechnol.* **16**, 27–31 (1998).
27. Chen, Y., Dougherty, E.R. & Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* **2**, 364–374 (1997).
28. Ermolaeva, O. *et al.* Data management and analysis for gene expression arrays. *Nature Genet.* **20**, 19–23 (1998).
29. Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
30. Debouck, C. & Goodfellow, P. DNA microarrays in drug discovery and development. *Nature Genet.* **21**, 48–50 (1999).
31. McAdams, H.H. & Shapiro, L. Circuit simulation of genetic networks. *Science* **269**, 650–656 (1995).
32. Yuh, C.H., Bolouri, H. & Davidson, E.H. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279**, 1896–1902 (1998).
33. Evan, G. & Littlewood, T. A matter of life and cell death. *Science* **281**, 1317–1322 (1998).

© 1999 Nature America Inc. • http://genetics.nature.com

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

THE POLYMERASE CHAIN REACTION (PCR)

1. IN VITRO DNA CLONING BY PCR

The polymerase chain reaction (PCR) consists of a rapid procedure for the in vitro enzymatic amplification of a specific segment of DNA (Mullis and Faloona, 1987). This powerful technique is an invention of Kary Mullis who sensed the possibility in 1983, and for which he received the Robert Koch Prize, as well as the Nobel Prize in Chemistry 1993. Like in vivo molecular cloning, PCR has paved the way for a wide variety of previously unthinkable experiments.

The number of PCR applications seems to be infinite and is still growing, particularly in the field of clinical diagnostics. They include direct cloning from genomic DNA or cDNA, in vitro mutagenesis and DNA engineering, genetic fingerprinting of samples analyzed in the context of medico-legal diagnostics, tests for the presence of infectious agents, diagnosis of genetic diseases, analysis of variations of allelic sequences, analysis of the structure of RNA transcripts, genomic footprinting, and direct nucleotide sequencing starting from genomic DNA and cDNA. Since the applications are such as numerous, it is desirable to achieve optimization conditions of this technique, which guarantee the achievement of consistent yields.
PCR is the most commonly used method in molecular biology, thanks to its particular **sensitivity**, **specificity** and operational **simplicity**. The correct application of this powerful diagnostic tool requires a solid knowledge of the basic principles of nucleic acid manipulation procedures, and of the possible drawbacks and dangers encountered during the analytical phases that lead to the result using the PCR technique.

In order to carry out a PCR reaction, a small amount of target DNA is added to a buffer solution containing the enzyme DNA polymerase, two appropriate oligonucleotide primers (called primers), the four deoxynucleotides triphosphates (dNTPs) that make up the DNA and the $MgCl_2$ cofactor (Magnesium Chloride). The primers are designed to delimit the target region to be amplified, so the ends of the sequence must be known in sufficient detail. Each primer binds to one of the two DNA strands so that their 3´ ends are directed towards the centre of the region to be amplified. In the case of transcribed gene sequences, the strand whose sequence is equal to the one of RNA will bind to the "antisense" DNA strand (or "template strand"). On the other hand, the "antisense primer" will have a complementary sequence, will have a sequence complementary to the one of the mRNA and will bind to the DNA strand "sense" (or "coding strand").

The PCR procedure has undergone automation, thanks to two important innovations. The original PCR procedure involved the use of the Klenow fragment of *E. coli* DNA polymerase I, thermolabile, with optimal activity at a temperature of 37°C, which underwent degradation during the denaturation phases of PCR. The discovery

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

of **thermostable polymerases** resistant to repeated denaturation phases and with an optimal activity at 72°C (*Taq* DNA polymerase; *Tth* DNA polymerase; *Pwo* DNA polymerase; *Vent* DNA polymerase) eliminated the needing for the addition of an aliquot of the enzyme to the reaction mixture after each stage of amplification. The most used is Taq DNA polymerase, isolated from the bacterium Thermus aquaticus, capable of withstanding very high temperatures (97° C) and capable of operating up to temperatures of 65°C.

It should be kept in mind that natural polymerases have typically different **activities** associates in the same enzyme:
5′-3′ DNA-dependent DNA polymerase, responsible for the template-dependent synthesis of a complementary DNA strand;
5′-3′ DNA exonuclease, able to remove DNA strands paired to the template in order to proceed with the DNA synthesis;
3′-5′ DNA exonuclease, which allows a step back to remove a mismatched nucleotide wrongly incorporated and to restore the correct one. This activity is absent in Taq polymerase but is present in the DNA polymerases found in some bacteria, such as *Pyrococcus furiosus* (*Pfu* polymerase). *Pfu* polymerase can thus increase the fidelity of the amplicon sequence, at the cost of a reduced yield of the product.

Besides, PCR is performed in machines (**Thermal Cycler**) where a computer checks the repeated changes in temperature and the duration of the respective phases.
Once prepared, the mixture is subjected to various cycles, through temperatures that allow the denaturation of the double-stranded DNA (91-96°C), the pairing of the primers with the target (50-65°C) and the synthesis of DNA by the polymerase (72°C) in order to amplify a product of predefined size and sequence exponentially. In addition to ensuring the automation of PCR, Taq polymerase has also led to other advantages. In fact, this enzyme, compared to non-thermostable polymerases, allows the use of higher temperatures in the pairing and extension phases, leading to an increase of the stringency during the hybridization between primer and template and therefore of the *specificity* of the amplification product. This circumstance results in an increase in the yield in the desired product.

The amplification reaction involves three phases:
(Animation of PCR at https://dnalc.cshl.edu/resources/animations/pcr.html)

1. Denaturation: the double-stranded DNA is denatured at a temperature of about 95°C, i.e. the two strands are separated, and it is converted into single-chain DNA.
2. Pairing (annealing): the oligonucleotide primers complementary to the two sides of the sequence to be amplified hybridize with the two denatured filaments, at a temperature that is approximately 2-5°C lower than the melting temperature (T_m) of the

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

primers themselves; their sequence is oriented to be able to guide DNA polymerization in the 5´-3´ direction in the stretch between the two regions to which they associate. The first part of this phase, during which the primers explore the entire starting DNA looking for the homologous sequences with which to pair, is often called the screening phase.
3. Extension: this phase, at a temperature higher than the previous one, involves anchoring the DNA polymerase to the site where the primer is bound, after which the enzyme extends the target DNA strand starting from the 3´ end made available from the primer and using the free dNTPs in solution. The primers are extended each in the direction of the other but on two different complementary chains leading to the synthesis of two double-stranded DNA molecules, copies of the target region delimited by the primers.

The first cycle is characterized by products of indeterminate length which tend to accumulate linearly with each subsequent cycle, i.e. the quantity present will be linearly proportional to the number of cycles carried out. From the second cycle onwards the first "short products" are produced, i.e. delimited by the ends 5´ of the two primers, the growth of which assumes an exponential trend with each subsequent amplification cycle. This growth can lead to a few million times amplification starting from the discrete fragment over 20-30 cycles.

The crucial chemical variable is the net synthesis of the product during the various cycles; due to this synthesis, the molecular balance between product, template, DNA polymerase, primer and deoxynucleotides changes with each cycle. As the amplification product accumulates, all the enzyme present is engaged during the extension phase, and the relationship between primer and template decreases, promoting the self-annealing of the DNA strand. Furthermore, since the size of the amplification product is much larger than that of the primers, such annealing can already occur at temperatures much higher than those of pairing the primers with the template. This fact will inevitably tend to occur during the cooling of the reaction mixture after the denaturation phase. When this reappearance becomes significant or *when the quantity of enzyme is limiting*, the reaction reaches saturation, and therefore exponential growth ceases; thus, the **plateau** phase was reached, and no more product is synthesized.

2. THE PCR PARAMETERS

Many important variables can affect the outcome of PCR.

Primers. The **oligonucleotides** that act as primers must be designed following well-defined criteria, which guarantee maximum pairing efficiency with the template DNA. The two members of the primer pair may be called in different ways:

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Sense primer
Direct primer, Forward primer
Left primer
"Bottom" primer
Primer 1

Antisense primer
Reverse primer, Back primer
Right primer
"Top" primer
Primer 2

The two regions on the template DNA that will be the target for the primers should not be far more than 1.5 kb, or DNA polymerase could not efficiently perform the synthesis. Maintaining the amplification product (**amplicon**) **size** below 700 bp ensures the possibility to obtain its whole sequence by Sanger standard sequencing.
When dealing with amplification from **RNA** (following its retrotranscription in DNA; RT-PCR) it is critical that the two primers are designed to anneal to **two different exons**, to clearly distinguish amplicons derived from RNA (with the size expected following the removal of the intervening intron sequence) rather than by contaminating DNA

"Biochemical" parameters
Any possible complementarity between the "forward" and "reverse" primer sequence must be avoided, in particular at the 3′ end, because it could cause the formation of a considerable quantity of **primer dimers** which reduce the yield of the desired product; primers with palindromic sequences and with extended secondary structures (due to self-complementarity) must also be avoided.
It is advisable to avoid the T or A bases at the 3′ end because they give a less stable pairing at the level of the critical site for the beginning of the extension. "**GC Clamp**" refers to the presence or one or more G or C bases at the 3′ end of the primer.
The dissociation **temperature** of the two primers, which basically depends on their length and concentration in C and G, must be roughly the same. From the structural point of view, an ideal primer must have a length between 18 and 28 nucleotides and composition in G+C (**GC content**) between 50% and 60%. These conditions guarantee a T_m between 50 and 80°C according to the simplified formula for determining the T_m starting from the related content of the pairs G/C or A/T:
Tm = 4 (G + C) + 2 (A + T)
As for the **concentration** of primers, this must be the same for both, and 0.1-0.4 µM is recommended for each (2.5-10 pmol in 25 µL of the reaction mixture). This concentration ensures that the excess of primer with respect to the template remains essentially constant. Higher concentrations are not necessary, such as 3 µM provided in previous protocols, and they are also disadvantageous as they lead to the formation of

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

primer-dimers and the increased probability of pairing errors.

"Biological" parameters

A fundamental characteristic of the primers must be the **uniqueness** of their sequence with respect to a complex eukaryotic genome, i.e. the guarantee that the probability of hybridising with sequences other than the desired one is extremely low. This prerogative is generally guaranteed by their length; in fact, a size of 18 or more nucleotides usually ensures the uniqueness of the sequence with respect to the genome.
Repeated sequences, or consecutive runs of the same base, are to be avoided.
In any case, this parameter must be checked within the database, checking in particular that it has not included in the oligonucleotide sequences homologous to Alu or mitochondrial DNA. Nowadays, numerous DNA sequence analysis software includes menus for primer design, also available in the web (Primer-BLAST).

Magnesium Chloride. One of the key variables of PCR is the concentration of Mg^{++} ions, usually variable between 0.5 and 2.5 mM, which has an important role both in terms of yield and specificity. It affects the reaction differently at high and low concentrations: too high concentrations of Mg^{++} *stabilise* the double-stranded DNA and hinder the complete denaturation of the product at each cycle, reducing the yield. Furthermore, the excess of Mg^{++} could also stabilise incorrect pairing between primer and template DNA, leading to an accumulation of unwanted product and therefore to low specificity. On the other hand, too low concentrations of magnesium (less than 0.5 mM) affect the extension phase, since Mg^{++} is required as a *co-factor* for the enzymatic activity of most DNA polymerases. Some Mg^{++} ions form complexes soluble with *dNTPs* in the reaction mixture. So there must be an ionic concentration for which there are optimal conditions in terms of yield and specificity.

Buffer. There are numerous buffers used in PCR, but the choice must be made according to the reaction conditions: characteristics of the target DNA and primers, as well as the reaction cycle. A 50 mM Tris-HCl buffer with a pH between 8.3 and 9 at 25°C is generally used (the pH of the Tris decreases with increasing temperature). To facilitate the pairing of the primers to the denatured DNA, up to 50 mM of KCl can be added to the reaction mixture. At concentrations above 50 mM, KCl, however, can be a Taq polymerase inhibitor.

Deoxynucleotide triphosphates (dNTPs). In most technical manuals, a concentration of dNTPs of 200 µM each is recommended, which guarantees an excess condition for the duration of the reaction. This concentration is sufficient for the stability of the dNTPs during repeated PCR cycles is such that approximately 50% of them remain after 50 cycles. dNTPs act as chelators towards the magnesium ions, changing their optimal concentration. Therefore, a quantity of dNTPs greater than 200 µM leads to an increase in the incidence of errors by the polymerase; millimolar concentrations of dNTPs completely inhibit the enzyme. Low concentrations of deoxynucleotides minimise phenomena of incorrect pairing towards sites other than the target, leading to advantages

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

in terms of specificity and accuracy (*fidelity*).

It is particularly important that the concentrations of the four deoxynucleotides are equivalent to prevent incorporation errors by the polymerase.

Polymerase. Thermostable DNA polymerases, like the other DNA polymerases, catalyze the synthesis of DNA from triphosphate nucleotides from a primer that has a 3´ free hydroxyl group. In general, they have maximum catalytic activity between 75°C and 80°C and activity substantially reduced at lower temperatures due to the change in pH. At 37°C, Taq polymerase has only about 10% of its maximum activity. Taq polymerase has a half-life that progressively decreases with increasing temperature (greater than 2 hours at 92.5°C, 40 minutes at 95°C and 5 minutes at 97.5°C).
Polymerases that lack 3´-5´ exonuclease activity generally have higher error rates than enzymes that exhibit this activity. The total error rate of the Taq polymerase is reported within a range between 1×10^4 and 2×10^5 errors per pair of bases.
As for the quantity, an excess of the enzyme could synthesize DNA from spurious interactions between primer and template. The Taq concentration suggested in most protocols is 0.5 unit (U) per 25 µL of reaction. This limiting concentration is required to control the specificity of the amplification reaction. Concentrations higher than 2.5 nM (1.25 U for 25 µL of reaction) must be avoided because they lead to a higher efficiency of the PCR only up to a certain point. In reality, such a high amount of enzyme can increase the yield in non-specific products, at the expense of the product of interest. In addition, the cost of the enzyme should not be underestimated.

DNA template. All the parameters discussed so far consider a total human genome quantity of 125 ng per 25 µL of reaction. The accuracy of the amplification depends on the DNA polymerase error rate, but it also depends on the number of initial copies of the template DNA.
It is not necessary for the sequence to be synthesized enzymatically to be present initially in pure form; it can also represent a small fraction in a complex mixture, such as the segment of a single copy gene in the total human genome. The sequence to be synthesized may initially be present as a discrete molecule or may be part of a larger molecule. In both cases, the reaction product will be a discrete double-stranded DNA molecule with ends corresponding to the 5´ end of the oligomers used.
Even relatively degraded DNA preparations can serve as useful templates for generating products of moderate length. The most important conditions concern purity and quantity. A large number of contaminants found in DNA preparations can reduce PCR efficiency. These include heme group, urea, SDS detergent, sodium acetate and, sometimes, residual components of purification from agarose gel.
Of course, the amount of template DNA must be sufficient to ensure the display of the PCR product using ethidium bromide (EtBr). Usually, 100 ng of genomic DNA is sufficient to reveal a PCR product from a single copy mammalian gene. Using too much template DNA might reduce the efficiency of the PCR reaction, both because $MgCl_2$ and other parameters are already optimized, and because of the greater quantity of contaminants that are introduced into the reaction mixture.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

The profile of the amplification cycles is also essential for carrying out the reaction.
Number of cycles. 30 amplification cycles are recommended for titration, but in some cases, if the yield is very poor, additional cycles are required. The number of cycles will depend on the molar ratio between the primer and the template. For a ratio of 10^6-10^7, 30 cycles should be adequate to generate sufficient material for viewing on ethidium bromide gel. For ratios higher or lower by two orders of magnitude, this value is respectively increased or decreased by 10 cycles.
Denaturation phase. For target sequences of 1 Kb or less, denature for 1 min at 94°C. For larger fragments, about one minute is added to the denaturation time for each additional kilobase. A long denaturation phase is particularly important at the beginning of the PCR, to ensure total denaturation of the starting DNA. For regions rich in G and C of the genomic DNA template, this phase is essential.
Annealing phase. It is critical for the specificity of amplification. A general scheme provides for the use of a pairing temperature (T_a) about 2-5°C lower than the T_m of the two primers used. A consequence of the use of a too low pairing temperature is that one or both primers can hybridize with sequences other than the target sequences (***non-specific amplification***).

This fact causes a drop in the yield of the desired product; on the contrary, a higher temperature leads to a reduction in the hybridization of the primers on the target DNA and therefore in the yield. The suggested time for pairing is 1 min. However, it can be reduced to 30 sec in the case of small reaction volumes (10-25 µL) that quickly reach the reaction equilibrium at low temperatures.
It is possible to use the same temperature profile to amplify different loci, even if the calculated temperature for pairing can be different. Templates for which the calculated temperature for pairing is higher than the standard profile may require a lower concentration of Mg ions. On the contrary, those with higher pairing temperatures may require a greater quantity of Mg ions to compensate for the too stringent temperature.
Extension phase. Taq polymerase is highly processive and extends 2-4 Kb per min (35-70 bases per sec). However, this parameter is strongly affected by the buffer pH, the saline concentration of the medium, as well as the nature of the target DNA. In general, for targets of size greater than 1 Kb, 1 min is sufficient for the extension phase. In reality, the rise time from the pairing temperature to the denaturation temperature is generally long enough to allow the complete synthesis of a target of 500 bp. For fragments greater than 1 Kb, about 1 min should be added to the extension phase for each additional Kb.
It is important to note that for 20 base primers, with a GC content of 60% or higher, T_m is in the range of temperatures at which the enzyme performs an efficient extension. Therefore, in this condition, the annealing and polymerization phases can be combined in a single step (*annealing-extension*).

3. PCR EFFICIENCY

The combination of all the parameters discussed so far determines, ultimately, what is the efficiency of the PCR reaction. A basic equation of the polymerase chain reaction describes the trend of amplification (Mullis, 1991):

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

$Yn = (1+ e)^n$
where Yn is the amplification factor after n cycles (although, for greater precision n should be considered the number of cycles minus one, given that the effective presence of the "short fragments" of interest begins, as was previously discussed, only starting from the second cycle) and e represents the efficiency of the reaction and assumes values between 0 and 1.
For example, if the starting DNA is made up of 3×10^5 (henceforth the 3E5 notation will be used) copies of the human genome, the observation of a band on standard gel would require 5E12 molecules of 100 bp length in an aliquot of 10 µL (Mullis, 1991). A typical PCR could foresee a starting DNA quantity equal to 300,000 copies of the human genome in a volume of 100 µL and could aim to amplify a 100 bp long fragment of a single copy human gene. With the formula described above, it is therefore possible to calculate the number of cycles necessary to amplify the desired segment, to obtain a visible band on gel:
$Yn = 5 \times 10^{11} /300,000 = 1.76 \times 10^{11}$
So, assuming optimistically equal to 1 the efficiency, we obtain that:
n = 20.7 cycles

If we assume that the reaction took place perfectly, in the last cycle 2.5×10^{11} DNA molecules would be amplified, which would, therefore, require 2.5×10^{11} Taq molecules, which correspond to about 10 units.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

THE EXPRESSED SEQUENCE TAGS (ESTs)

The method for rapidly obtaining **partial mRNA sequences**, published by Adams et al. in 1991 and described below, gave a tremendous and sudden acceleration to the rhythm of identification and cloning of mRNA, to the point that during the 1990s the cloning in the form of cDNA of about 1,000 new human mRNAs / year was reported in the literature. In this way, the characterisation of the entire series of the approximately 20,000 cDNAs corresponding to human RNAs coding for proteins was substantially completed in the first years after 2000.

The fundamental idea to arrive at a faster identification of the transcribed sequences was to create databases containing partial sequences derived from RNA. To this end, the **RNA** extracted from various tissues is converted into **cDNA** with the use of reverse transcriptase and cloned according to the cDNA library method. While until then the typical approach aimed to seek only a specific sequence of interest, present in a particular bacterial clone that had incorporated it, the new strategy was distinguished by the intensive **automation** of the analysis procedure, which allowed to randomly characterise a **large number of bacterial clones**, favouring speed over accuracy and completeness of sequencing (Figure, from Fantoni et al., "Genetica", Piccin).

In this way, it is possible to determine the sequence of **a few hundred bases located at one of the two ends of the cDNA insert,** which on the whole is usually about 1-3 kb long. The sequences obtained in this way are called **EST (expressed sequence tags)**, that is, partial RNA sequences. A single EST is, therefore, a short sequence located at one end of a specific cDNA, or more briefly represents a fragment of a specific RNA. ESTs are published in specialised databases (Boguski et al., 1993) and are publicly available for computer sequence analysis. Furthermore, representative cells of each bacterial colony are set aside by the robots and stored at low temperature to allow the replication of the cloned cDNA as desired and its complete characterisation, if its sequence is of interest.

It is therefore assumed by definition that, *if a particular sequence of bases is found in the EST database, it belongs to a transcript*. It is also possible to "assemble" by the computer the fragments of EST sequences referable to the same transcript by exploiting the existence of regions of coincident sequence in the various fragments obtained at random. In this way, it is possible to quickly identify, by only computer processing, genome sequences that have been transcribed (genes) without having to deal with the "background" of extragenic sequences, as happens in traditional "genome projects".

Since different RNAs are present in the cells depending on the tissue to which they belong and the conditions in which they are found, dozens of cDNA libraries are prepared in the larger EST projects, each starting from RNA extracted from a **specific tissue or organ** which is in a given physiological (e.g., stage of development) or pathological (e.g., neoplastic proliferation) condition.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

The availability of the complete sequence of the human genome and the development of powerful bioinformatic analysis tools make it possible today to determine the position of the EST on the chromosome with the simple comparison conducted on the computer between an EST sequence and the complete sequence genomic DNA.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Difficulty in obtaining the complete mRNA coding sequence at 5' region (5' end mRNA artifact): Causes, consequences in biology and medicine and possible solutions for obtaining the actual amino acid sequence of proteins (Review)

LORENZA VITALE[1], MARIA CARACAUSI[1], RAFFAELLA CASADEI[2], MARIA CHIARA PELLERI[1] and ALLISON PIOVESAN[1]

[1]Department of Experimental, Diagnostic and Specialty Medicine (DIMES), Unit of Histology, Embryology and Applied Biology, University of Bologna, I-40126 Bologna; [2]Department for Life Quality Studies, University of Bologna, I-47921 Rimini, Italy

Received November 22, 2016; Accepted March 16, 2017

DOI: 10.3892/ijmm.2017.2942

Abstract. The known difficulty in obtaining the actual full length, complete sequence of a messenger RNA (mRNA) may lead to the erroneous determination of its coding sequence at the 5' region (5' end mRNA artifact), and consequently to the wrong assignment of the translation start codon, leading to the inaccurate prediction of the encoded polypeptide at its amino terminus. Among the known human genes whose study was affected by this artifact, we can include disco interacting protein 2 homolog A (*DIP2A*; *KIAA0184*), Down syndrome critical region 1 (*DSCR1*), SON DNA binding protein (*SON*), trefoil factor 3 (*TFF3*) and URB1 ribosome biogenesis 1 homolog (*URB1*; *KIAA0539*) on chromosome 21, as well as receptor for activated C kinase 1 (*RACK1*, also known as *GNB2L1*), glutaminyl-tRNA synthetase (*QARS*) and tyrosyl-DNA phosphodiesterase 2 (*TDP2*) along with another 474 loci, including interleukin 16 (*IL16*). In this review, we discuss the causes of this issue, its quantitative incidence in biomedical research, the consequences in biology and medicine, and the possible solutions for obtaining the actual amino acid sequence of proteins in the post-genomics era.

Contents

Correspondence to: Dr Maria Chiara Pelleri, Department of Experimental, Diagnostic and Specialty Medicine (DIMES), Unit of Histology, Embryology and Applied Biology, University of Bologna, Via Belmeloro 8, I-40126 Bologna (BO), Italy
E-mail: mariachiara.pelleri2@unibo.it

Key words: messenger RNA 5' region, full-length cDNA, coding sequence, protein prediction, protein sequence

1. Introduction

Since the late 1990s, the availability of public, large databases containing growing information about genes, gene products (RNAs and proteins), genomes and molecular functions has radically changed the traditional approach to gene discovery and characterization. Combining the deposited data about informational molecules (1,2) obtained from multiple species is a straightforward method to gain rapid knowledge about the structure of an organism's genes and gene products, which in turn may be used to obtain clues as to the function of each individual gene. While this possibility has allowed the generation of an amount of data incomparable to what was obtained by classic molecular biology methods used in the pre-genomic era (3), the fact that the quality and degree of the information available for an individual gene may tend to decrease is less evident. For example, if we consider the characterization of the messenger RNA (mRNA) expressed by a human locus, through the 1980s and 1990s it was typical to obtain accurate information about the total mRNA size and tissue distribution by northern blot analysis and about the transcription initiation sites by S1 nuclease mapping, primer extension and run-off assays (4). In later years, mRNA full-length sequences were obtained by tailored experiments designed for polymerase chain reaction (PCR) amplification of DNA complementary to RNA (cDNA) ends [rapid amplification of cDNA ends (RACE)], alternative splicing information by cDNA *in vivo* and *in vitro* cloning of individual RNA isoforms, and protein sequences by *in vitro* translation and polypeptide biochemical analysis. Indeed, genes were usually studied on a

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

one-by-one basis, and there was the possibility to cross-check data made available through different methods (5). An example would be the comparison of the mRNA length deduced from northern blotting (taking into account the polyadenylated tail) and the one of the isolated cDNA (6), or the comparison of the molecular weight of a known protein (7) and the one of the polypeptide predicted to be encoded by the open reading frame (ORF)/coding sequence (CDS) of its relative cDNA.

New large-scale methods cannot always reach the resolution of previous ones; therefore, while they set a new standard in the methods used in genetics, more detailed analysis aimed at characterizing each individual gene remains necessary in order to avoid incomplete or erroneous knowledge of the gene structure and function. However, the genome-scale information has been in turn invaluable in effectively directing further investigations needed for each genomic locus using classical methods. This has been shown in particular for the human genome, by a large corpus of millions of short sequences (a few hundred base pairs in length) which has been derived by partial, single-pass sequencing of the cDNA clones from RNA of specific tissues (8). These have accumulated in the expressed sequence tag (EST) database since its creation >20 years ago (9). A variety of EST-based methods (10,11) were then used for the rapid *in silico* cloning of genes (12), determining differential gene expression (13), characterizing alternative forms of transcripts derived from alternative splicing (14,15), and defining at least one complete ORF (16) for each mRNA. This last point is a well-known issue in molecular biology and genomics, with relevant consequences for the prediction of the gene product structure and function, and will be analyzed in detail in this review.

2. The 5' end mRNA artifact

According to the classic molecular biology central dogma, the final effector of the genetic information is the protein (a chain of amino acids) encoded from a given gene; thus it is crucial to know the basic, primary structure of the protein (its amino acid sequence). A landmark in this field was the sequencing of the two amino acid chains composing human insulin by Sanger (17). Polypeptide sequencing has the advantage of determining the natural primary structure of the polypeptide chain, and in particular the actual first amino acid of the sequence, thanks to the ability of fluorodinitrobenzene to react with the N-terminal amino group at one end of the chain. Key subsequent advancements were the recognition that, due to the colinearity of nucleic acids and proteins and to the mechanisms of mRNA translation, amino terminal amino acids are encoded by the 5' end of the mRNA (18). Therefore, when Sanger *et al* proposed a new effective method to sequence DNA (19), it became evident that it was much more convenient to sequence the nucleic acids rather than the proteins, and that the amino acid sequence of gene products could be conveniently deduced from the nucleotide sequence of the relative cloned cDNA. This change of experimental paradigm led to 'reverse genetics' (20), the passage from nucleic acid sequences to their functions rather than the contrary as in classic genetics and has had the fundamental consequence that actually, since the late 1970s, the vast majority of protein sequences were no longer directly determined, but were predicted following sequencing of the relative cDNAs according to rules for recognition of the start codon (first-AUG rule, optimal sequence context) and the genetic code (21).

While this advancement greatly sped up the pace of the availability of protein sequences, it should be kept in mind that all standard experimental methods for the cloning of cDNA are affected by a potential inability to effectively clone the 5' region of mRNA in its completeness (22). This is due to the reverse transcriptase (RT) failure to extend first-strand cDNA along the full length of the mRNA template toward its 5' end (22) (Fig. 1), an operation whose success depends on the natural processivity of the enzyme, as well as its quality, the integrity of the RNA, the secondary structures assumed by the 5' region of the mRNA hampering the RT progression and the reaction conditions (23).

It should be highlighted here that, due to the intrinsic functional mechanisms of the polymerases able to generate DNA copies of mRNAs, cDNA is typically obtained through a primer starting polymerization from the 3' region of the mRNA [e.g., a poly(dT) oligonucleotide pairing with the poly(dA) tail present in the vast majority of mRNAs]. This implies that a cDNA collection is by definition enriched in the 3' regions of the mRNAs, and consequently it is expected that the prediction of the amino acid sequence at the carboxy terminus of the gene product is more accurate than the one at the amino terminus. This problem was recognized early on, in the publication of the first sequenced human cDNA, the one for the β chain of hemoglobin in 1977 when the 5'-untranslated region (UTR) was the last region to be reported in December (24) following previous descriptions of 3'-UTR in April (25) and CDS in July (26): 'cloning cDNA has proven to be a most valuable technique for sequencing mRNA (27,28). During the construction of double-stranded cDNA, however, a considerable number of 5'-non-coding region sequences are lost. The independent sequencing of this region will therefore be a necessary step to complete our knowledge of the primary structure of any mRNA' (24); Okayama and Berg clearly wrote in 1982: 'obtaining cloned cDNAs with complete 5'-UTR and protein-coding sequences is rare, particularly if the mRNA codes for a large protein. Although such truncated cDNAs are still useful as hybridization probes, they cannot direct the synthesis of complete proteins after their introduction into bacterial or mammalian cells via appropriate expression vectors' (23).

A flourishing of reports in the 1980s presented the determination of the often called 'cDNA full-length sequence' for many human genes. For the reasons discussed, the concept of the 'full-length sequence' becomes *de facto* equivalent to the one of 'completeness of mRNA sequence at its 5' end' and remains an open issue in molecular biology as cDNA incompletely representing the 5' end of the relative mRNA may lead to the incorrect assignment of the first AUG codon. In these cases, should an additional upstream AUG - in frame with the previously determined one - have been identified in a more complete mRNA 5' end, it would have been considered the actual translation start codon, thus extending the predicted amino terminus sequence of the product. Assignment of the inexact start codon leads to a series of subsequent relevant errors in the experimental study of the relative cDNA. We therefore introduced the term '5' end mRNA artifact' to refer

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

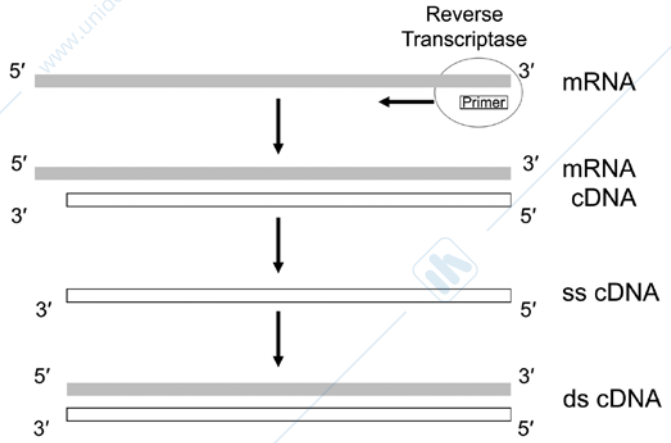


Figure 1. The 5' end mRNA artifact. cDNA is typically obtained through a primer starting polymerization from the 3' region of the mRNA by reverse transcriptase. The natural processivity of the enzyme, as well as its quality, the integrity of the RNA and the secondary structures assumed by the 5' region of the mRNA may hamper the reverse transcriptase progression, causing a failure in the polymerization of the first-strand cDNA along the full length of the mRNA template toward its 5' end, affecting all further experiments, including the assignment of the first AUG codon. ss, single-stranded; ds, double-stranded.

to the incorrect assignment of the first translation codon (AUG sequence) in an mRNA, due to the incomplete determination of its 5' end sequence (29).

From the experimental point of view, the recognition of this technical issue, although often without systematic investigation of its possible consequences for genome annotation, has led to the development of several methods to determine the full length mRNA sequence on a large scale. Some were based on the presence of the 'cap' at the true 5' end of the mRNA [reviewed in (30)], such as 5' cap trapping (31) and cap analysis of gene expression (CAGE) (32). Systematic empirical annotation of a set of transcript products by 5' RACE (33) has also been employed, as well as after the introduction of microarray-based platforms, hybridization of RNA on high-density resolution tiling arrays (34). However, these techniques were found to be experimentally labor-intensive and they have not been routinely applied.

Concurrently, the growing incorporation of information derived from individual cDNA and large-scale sequencing projects, including those specifically designed to characterize mRNA 5' end (31,35,36), led to a continuous refinement and improvement of completeness at the 5' region of deposited and verified mRNA reference sequences (e.g., RefSeq, https://www.ncbi.nlm.nih.gov/refseq/), as also regarding the corresponding protein-coding sequences. Therefore, it became possible to exploit the data from EST or other large-scale RNA sequencing projects to verify if sequence analysis could be optimized to reveal the extension of the 5' region of known mRNAs and possibly the consequential redefinition of the amino acid sequence of the encoded products.

The recent availability of massive RNA-sequencing (RNA-Seq) methods for the generation of transcriptome sequence databases (37) offers a new potential tool to deal with the issue, although to date it appears not to have been systematically used to this aim. Moreover, information about sequences possibly extending the knowledge of the 5' end of mRNA is not easily derivable from RNA-Seq data in comparison with the EST-based approach, due to short sequence reads typically obtained by this technique, as well as difficulty in building full-length transcript structures.

Furthermore, a ribosome footprinting profiling strategy based upon high-throughput sequencing of ribosome-protected mRNA fragments has been developed, enabling the genome-wide investigation of translation (38). This technique, used in combination with initiation-specific translation inhibitors, allows the identification of translation initiation with subcodon or even single-nucleotide resolution and was successfully exploited in order to predict also additional upstream AUG codons (39-41).

Finally, we should note the existence of ORFs and out-of-frame AUGs located in the 5'-UTR, upstream of the main coding region (42). These situations are different from the artifact reported herein as they do not extend the known coding region, but are implicated in the regulation of gene expression by modulating mRNA stability and translation (42,43).

3. Systematic identification of incomplete 5' end region in human known mRNAs

The theoretical possibility that the presence of a more precise knowledge of the mRNA 5' end sequence may lead to consequential correction of the previously accepted predicted product appeared in several reports in the form of anecdotal evidence randomly found for single genes that were under detailed investigation. For example, mRNA CDS was extended in this way for *RANBP9/RanBPM* gene (RAN binding protein 9, on 6p23), where the study performed by Nishitani *et al* (44) allowed the addition of 230 new amino acids. In the case of nuclear factor, erythroid 2-like 3 (*NFE2L3*) gene (on 7p15.2), the corresponding #AB010812.1 mRNA sequence of 2,174 bp in length derived from Kobayashi *et al* (45) was replaced by the sequence #AF134891.1 of 2,618 bp, leading to the addition of 294 new amino acids to the predicted protein. The study performed by Nomura *et al* (46) for *SP2* gene (Sp2 transcription factor, on 17q21.32) allowed the release of the #D28588.1 mRNA sequence entry recording a CDS of 3,288 bp leading to the addition of 111 new amino acids compared to the previous #M97190 entry of 2,063 bp provided by Kingsley and Winoto (47). The coding nature of these extensions was also supported by very high similarity with the respective murine hortologs (29). These and other similar reports suggested that a high-throughput approach was desirable to discover all the incompletenesses in the CDSs (Table I).

Regarding our group, as a first approach to the issue, due to our interest in an integrated route to identifying new pathogenesis-based therapeutic approaches for trisomy 21 (Down syndrome) (48,49), we focused on the known, well-characterized genes present in the original map of human chromosome 21 (Hsa21), manually analyzing 109 RefSeq mRNA sequences catalogued as 'category: known' by Hattori *et al* (50), and linked to at least one published report, for the presence of an in-frame stop codon upstream of the described ATG. In 49 cases, the finding of such a stop codon allowed the exclusion of the possibility that the recorded 5'-UTR sequence may actually be part of a longer CDS (51). The sequence of the remaining 60 mRNAs

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Table I. Main published results of systematic search for completeness of mRNA 5' CDS region.

Ref.	Year	Organism	Method	mRNAs	Extended 5' CDS[a]
(35)	2000	*H. sapiens*	Oligo-capping	954	68 (7.1%)
(29)	2003	*H. sapiens*	Manual and automated sequence analysis	13,124	556 (4.2%)
(53)	2007	*D. rerio*	Automated sequence analysis	8,528	285 (3.3%)
(39)	2011	Mouse embrionic stem cells	Ribosome footprinting profiling and support vector machine (SVM)-based machine learning strategy	4,994	570 (11.4%)
(54)	2012	*H. sapiens*	Fully automated sequence analysis	18,665	477 (2.6%)
(40)	2012	*H. sapiens*	Ribosome footprinting profiling and neural network prediction	5,062	6 AUG (0.1%) and 540 non-AUG (10.7%)
(55)	2014	*M. musculus*	Fully automated sequence analysis	20,221	351 (1.7%)
(41)	2014	*H. sapiens*	Ribosome footprinting profiling and manual analysis	1,255	17 (1.4%)
		M. musculus	Ribosome footprinting profiling and manual analysis	930	4 AUG (0.4%) and 13 non-AUG (1.4%)

[a]Estimation. CDS, coding sequence; *H. sapiens*, *Homo sapiens*; *D. rerio*, *Danio rerio* (zebrafish); *M. musculus*, *Mus musculus* (mouse).

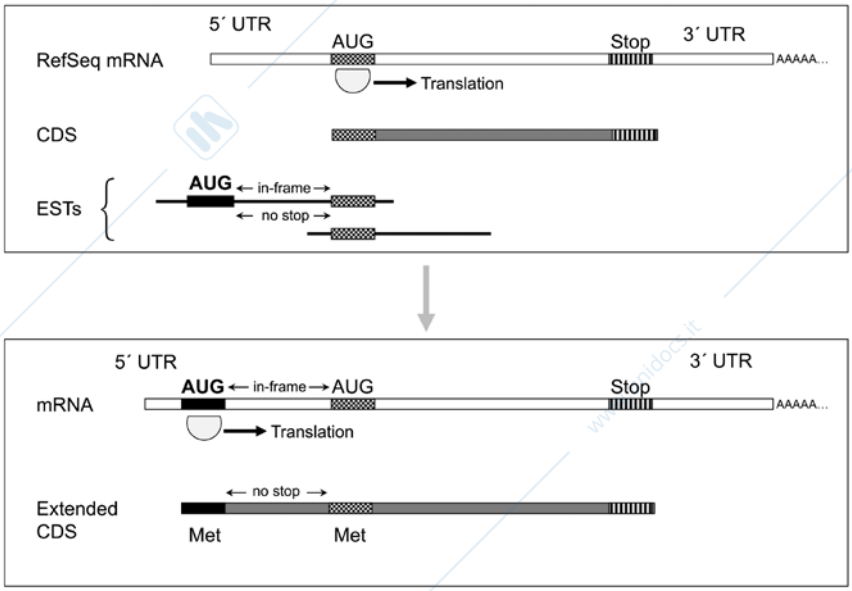


Figure 2. Identification and correction of incomplete 5' end regions. Possible EST sequence candidates for extending the known mRNA 5'-coding region are selected for the presence of an upstream in-frame AUG codon and absence of any stop codon between the previously known and the newly determined AUG codons. The upstream in-frame AUG codon becomes the actual translation start codon, thus encoding for a new Met and extending the predicted amino terminus sequence of the mRNA product. EST, expressed sequence tag; Met, methionine.

in which bases in the 5'-UTR could on the contrary be consistent with the presence of translated codons was systematically aligned with sequences available in databanks using Basic Local Alignment Search Tool (BLAST software, http://www.ncbi.nih.gov/BLAST/), leading to the discovery of a total of 20 genes for which EST (or also non-EST RNA sequences) homology suggested the existence of mRNAs more complete at 5' terminus. They putatively encode for protein products longer at their amino terminus, due to the presence of a previously unknown start codon in frame with and upstream of the described one (Fig. 2). Experimental evidence for the existence of these transcripts was finally obtained, following RT-PCR and sequencing, for five loci: down syndrome critical region 1 (*DSCR1*) [now regulator of calcineurin 1 (*RCAN1*)], disco interacting protein 2 homolog A (*DIP2A*; *KIAA0184*), URB1 ribosome biogenesis 1 homolog (*URB1; KIAA0539I*), SON DNA binding protein (*SON*) and trefoil factor 3 (*TFF3*) (29). In these cases, both of the following conditions occurred: an extension of described exon 1 predicted new coding codons upstream of

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

the known AUG; and a novel AUG was present upstream of these codons, in frame with the previously described AUG and without any intervening stop codon. This thus suggests that, following the rules of translation initiation [reviewed by Kozak (21)], the actual CDS should be considered as the one included between the novel 'first-AUG' and the known stop (Fig. 2). It was observed that no known mechanism hampers the possibility that the newly identified start codon is not the point of actual translation as the use of 'internal' AUGs, enabling additional initiation events at downstream AUG codons in some mRNAs may occur only in three well-defined circumstances (21): re-initiation, which does not apply to the mRNAs investigated by this approach, as the newly determined AUG is not part of a small upstream ORF separated from the main ORF by a stop codon; context-dependent leaky scanning, which may be excluded as we considered the concordance with the Kozak sequence (21,52) for the novel AUGs, observing full (sometimes better) compatibility with the use of the novel AUG (29); a third mechanism, that is the use of internal ribosome entry site (IRES) sequence modules, adopted only by some known viral mRNAs.

These positive results suggested to extend the approach to the whole set of human RefSeq mRNAs known at the time (n=13,124), following automation by a simple program to detect the presence or the absence of an in-frame stop in the described 5'-UTR of an mRNA. The percentage of the latter type of mRNA in the set (51%) was very similar to the one found for the Hsa21 gene set (55%), thus estimating that, in proportion, the CDS of 556 known human mRNAs might be incomplete at the 5' end (29).

This approach required manual curation to analyze in detail, by sequence comparison, any mRNA candidate to have an incomplete CDS at 5' region. An improvement of the algorithm was then published and applied with success to zebrafish [see below (53)], showing that the automated detection of putative additional bases at the known 5' end of a set of mRNAs following elaboration of multiple results of sequence comparison analysis (by BLAST tool) was possible. Some technical limitations of the used environment made the implementation of this pipeline difficult for the much more numerous human sequences which hampered progress in this direction for a while. Further improvement of the automated EST-based approach (5'_ORF_Extender 2.0, freely available at http://apollo11.isto.unibo.it/software/) finally made the systematic identification (Fig. 2) of CDSs at the 5' end of all human known mRNAs possible, parsing >7 million BLAT alignments and thus finding 477 human loci out of 18,665 analyzed (Table I), with an extension of their RNA 5' CDS identified in detail (54). In addition, in this study, a proof-of-concept confirmation was obtained by *in vitro* cloning and sequencing for some example genes: *GNB2L1* [now receptor for activated C kinase 1 (*RACK1*)], glutaminyl-tRNA synthetase (*QARS*) and tyrosyl-DNA phosphodiesterase 2 (*TDP2*) cDNAs. On the other hand, a list of 20,775 human mRNAs where the presence of an in-frame stop codon upstream of the known start codon indicates completeness of the CDS at 5' end in the current form was generated (54). This approach could also be aimed at the different 5'-UTR sequence identification, but the length of the bases aligned upstream of the novel AUG is usually too short to allow this type of investigation. In addition, should the length be long enough, the analysis would require an ad hoc algorithm able to discriminate mRNA isoforms of this type, including mapping of the newly determined 5'-UTR to the genome to derive the alternative transcription/splicing events responsible for the different 5'-UTR sequences.

While this review is more focused on human mRNAs for the possible repercussion in medicine, it should be noted that similar results are to be expected for the genomes of other organisms due to the sharing of common molecular techniques, whose limitations are at the basis of the artifact. Actually, studies on two of the most commonly used model organisms for the investigation of the human genome, *Danio rerio* (zebrafish) and *Mus musculus* (domestic mouse) have confirmed this expectation. A novel proposed automated approach (5'_ORF_Extender 1.0) was able to systematically compare available ESTs with all the zebrafish experimentally determined mRNA sequences, identify additional sequence stretches at 5' region and scan for the presence of all conditions needed to define a new, extended putative ORF. The tool identified 285 (3.3%) mRNAs with putatively incomplete ORFs at the 5' region and, in three example selected cases (*selt1a*, *unc119.2* and *nppa* or selenoprotein T 1a, unc-119 lipid binding chaperone B homolog 2 and natriuretic peptide A, respectively), the extended coding region at 5' end was experimentally demonstrated (53). As regards the mouse mRNAs, the application of the improved method used for human transcripts (54) showed that in 351 mouse loci, out of 20,221 analyzed, an extension of the mRNA 5'-coding region could be identified. Experimental confirmation was obtained by *in vitro* cloning and sequencing for adenomatosis polyposis coli 2 (*Apc2*) and MAP kinase-interacting serine/threonine kinase 2 (*Mknk2*) cDNAs and a list of 16,330 mouse mRNAs with estimated complete CDS at 5' end was provided (55). Remarkably, 82% of the results were original and have not been identified by the annotation pipelines used in the main mouse genome databases and genome browser (55). The diffusion of the 5' end mRNA artifact may thus be considered approximately constant from lower vertebrates to humans because the methods used to characterize the relative mRNAs are the same or very similar (Table I).

The identification of the most upstream definable start codon does not exclude that a downstream AUG codon may also be used by the ribosome, a phenomenon known as alternative translation (56). It has been shown that alternative translation start sites tend to be conserved in eukaryotic genomes, providing a functional mechanism under selection for increased efficiency of translation and/or for obtainment of different N-terminal protein variants (57). It has also already been noted that this type of analysis cannot formally exclude that the extended ORF may derive from alternative transcription starting site (due to alternative promoter usage) and/or splicing of the investigated locus (53). However, it reveals in any case that additional coding sequences not previously identified exist, as may be confirmed by phylogenetic comparison at the amino acid level (53). As in the case of any other computer prediction, further investigation is required, *in silico* but especially *in vitro*, for a fine characterization of the putative model.

While the published approaches have considered algorithms assuming that the start codon has an AUG sequence, it should be noted that in a minor percentage of mRNA CDSs the start codon may have alternative sequences, particularly CUG, UUG,

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Table II. Possible consequences of incomplete determination of mRNA 5' CDS region for example human genes.

	Symbol	Ref.	AAs[a]	Ref. 2
At protein level				
Errors in determining the 3D protein structure	ALDOC	(59)	87	(54)
Prediction of an incomplete polypeptide	QARS	(60)	18	(54)
Production of an incomplete polypeptide	IL16	(61)	47	(54)
Lack of description of functional protein domains	SON	http://www.ncbi.nlm.nih.gov/gene/6651	968	(29)
Errors in identifying protein localization	RANBP9/RanBPM	(63)	230	(44)
Failure to predict alternative polypeptides	UMOD	http://www.ncbi.nlm.nih.gov/gene/7369	49 or 28	(54)
Errors in identifying ortholog products	DSCR1.1	(66)	55	(29)
	Symbol	Ref.	nts[a]	Ref. 2
At cDNA level				
Failure to screen the complete CDS for mutations	*ADAR*	http://omim.org/entry/146920	48	(54)
Incomplete cDNA in two-hybrid test for function	*DSCR1*	(65)	55	(29)
Potential errors in designing morpholino oligos	*unc-119.2* (*Danio rerio*)	(77)	58	(53)
At gene structure level				
Failure to identify the full extension of the gene/ labeling of genic regions as intergenic space	*DIP2A*	(71)	82,895	(29)
Failure to identify actual promoter regions	*TFF3*	(72)	170	(29)

[a]AAs or nts added to the previously recorded protein or nucleic acid sequence, respectively, following the analysis cited as Ref. 2. CDS, coding sequence; AAs, amino acids; nts, nucleotides; ALDOC, Aldolase, Fructose-Bisphosphate C; QARS, glutaminyl-tRNA synthetase; IL16, interleukin 16; SON, SON DNA binding protein; RANBP9, RAN Binding protein 9; UMOD, uromodulin; DSCR1, down syndrome critical region 1; *ADAR*, adenosine deaminase, RNA specific; *DIP2A*, disco interacting protein 2 homolog A; *TFF3*, trefoil factor 3.

GUG, ACG, AUA and AUU (58). Actually, recent experiments have confirmed this phenomenon and suggested that it may be more frequent than was previously assumed. Therefore, when the use of a non-AUG codon is known or suspected, further analysis not included in standard pipelines should be performed in individual cases to identify in frame upstream non-AUG start codons which may also be responsible of encoding proteins longer that the ones previously described.

4. Consequences of 5' end mRNA artifact in biology and medicine

The 5' end mRNA artifact is expected, and demonstrated, to cause a chain of consequences in biomedical research, that will be now listed and discussed (Table II). The first obvious issue associated with the artifact is the negative consequence on the study of product structure and function (59). The possibility that vast amounts of studies are based on incorrect starting data is real. For instance, it occurred in the functional characterization of a polypeptide expressed from its predicted incomplete DNA (60) and in a functional study of the cytokine interleukin 16 (IL16) (61), where the product appears to be expressed from an incomplete cDNA (Table II).

The recording of protein sequences incomplete at their amino terminus in the genomic databases may also cause the failure to identify functionally remarkable protein domain sequences (Table II); in particular, sequences located at the amino terminus of proteins may be represented by signal peptide sequences directing delivery of the protein to its final destination (62,63) and may also affect its half-life (64).

In addition, there is the possibility to underestimate alternative splicing at the 5' terminus of genes and to not predict the corresponding alternative protein gene products (Table II). The statement in the classic article by Okayama and Berg still holds true: 'indeed, it was comparison between cloned cDNAs and their genomic counterparts that uncovered the existence of intervening sequences and splicing' (23). Moreover, the design of a mutation screening aimed at identifying pathological variations in the coding sequences could be affected by the incomplete knowledge of the CDS, a circumstance that could occasionally explain the failure to find expected mutations in candidate or established disease genes, and could possibly lead to inaccurate genotype/phenotype correlations (Table II). From a functional point of view, the new amino acid sequence could be responsible for new interactions. The possibility of designing molecules with pharmacological activity based on binding to proteins expressed as bait in a two-hybrid test from incomplete cDNAs (65) emphasizes the importance of knowing the actual primary structure of the protein. Finally, the presence of a truncated protein sequence in the genomic databases may also be at the origin of a chain of errors in the prediction of orthologs in other species. In particular, the genome annota-

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

tion pipelines will tend to propagate the truncated sequence in the predicted model proteins. For instance, the error in determination of the highly similar corresponding murine DSCR1 ortholog (66) underlines that a bias deriving from the original human incomplete data negatively affected the modeling of the murine DSCR1 product sequence.

Due to the complex structure of the human loci (67-70), errors in establishing an accurate cDNA sequence may also finally cause drawbacks in the study of genomic organization of a gene due to the tight connections between DNA and RNA (Table II). If a cDNA incomplete at its 5' terminus is used to establish the genomic structure of a locus, this could cause failure to recognize genomic sequences as part of the locus (71). As a secondary consequence, classification of a genic region as intergenic may keep the 'search space' for novel genes artificially expanded (71). Due to the physical proximity of the gene promoter region and the corresponding mRNA 5' region, a sequence supposed to be proximal to the transcription start site and annotated as promoter could be actually part of a longer mRNA, as was shown for *TFF3* (72,29). This issue may further increase the difficulty in identifying promoter sequences that do not have regular start and stop signals or characteristic cross-species conservations as the CDSs, and can even present with diverged sequences among distant species, while being functionally conserved (73). On the other hand, a non-exact delimitation between 5'-UTR and CDS could lead to errors in the knowledge of the 5'-UTR sequence itself and in the interpretation of its role in the control of translation (74). Although this last class of consequences does not directly affect the prediction of the CDS, they should be considered as a further incentive to not underestimate the relevance of this artifact due to the central role of the 5' terminus in gene expression regulation pathways. The knowledge of the true mRNA end is also useful in designing probes specific for this region that may be more variable between similar loci or isoforms from the same locus rather than the central, coding region. This is relevant regarding the possibility to extract from publicly available microarray datasets quantitative reference measures for the expression values of the whole complement of the genes of both normal (75) or pathologic (76) transcriptomes. Exact knowledge of mRNA 5' region also affects the choice of morpholino oligonucleotides, in particular in zebrafish (77), used in knockdown experiments (Table II).

The artifact may also be a source for errors in other types of genomic analysis, although in these cases the consequences are expected not to be relevant, as the alteration of calculations is likely to represent a small deviation, and not for immediate medical application of these analyses [e.g., estimation of codon usage at a genomic scale (78), although the knowledge of the whole set of codons in a cDNA could affect the technology of the production of the translated product in a host (79)].

5. Possible solutions for improving the knowledge of the 5'-coding regions in mRNAs

Several methods have been described with the aim of knowing with more precision the 5' mRNA end, thus excluding that its CDS may be incompletely predicted. The first were devised in the 1990s and were based on experimental protocols exploiting the capability of dedicated techniques to identify the first bases transcribed from DNA or the first bases following the cap in mature mRNAs. These methods have been cited in the Introduction section and remain valid, although they were often labor-intensive and not routinely used.

A second group of methods is based on computational biology approaches, with the advantage of providing a first systematic screening leading to exclusion of a relevant number of genes as candidates for the 5' end mRNA artifact. Due to the availability of throughput results of an EST-based approach of this type (54), it is advisable to perform a simple first check against these results for a gene of interest before assuming that the predicted product is the one recorded in the current version of databases. Continuous refinement over time of the human mRNA sequences has led to the current estimation of 259 nucleotides as the mean 5'-UTR size (80), so there is the concrete possibility that extended protein-coding sequences could actually be hidden in longer 5'-UTRs. Further developments of the computational analysis of high-throughput cDNA sequencing methods (RNA-Seq) should also provide a means to increase the characterization of whole sequences of human transcripts. Several studies have been performed to implement RNA-Seq methods of profiling mRNA 5' ends in *Drosophila melanogaster* (81,82).

Finally, recent developments of proteomics research open the way for a different, specular approach to the problem. Knowledge of protein sequences obtained by massive analysis of polypeptide nuclear magnetic resonance (NMR) or mass spectrometry (MS) spectra, in particular oriented to N-terminal sequencing (83,84), might be used for a reverse search for genomic sequences predicted to be translated in the corresponding identified protein sequences. This thus resembles the first protein-toward-DNA experimental flow but at on a genomic scale and largely based on computational methods.

In conclusion, we have presented evidence that current methods of genomics research are subject to a possible artifact regarding the exact determination of the mRNA 5' region sequence and the consequences that this may have on the annotation, as well as on the experimental study of both genes and gene products. While there are several strategies to deal with this issue, the most important issue appears to bring this possibility to the attention of the scientific community so that it is taken into account when planning experiments in molecular biology and genetics.

Acknowledgements

M.C.'s fellowship has been co-funded by a donation from Fondazione Umano Progresso (Milano, Italy) and by a grant from Fondazione del Monte di Bologna e Ravenna (Bologna, Italy). M.C.P.'s fellowship has been co-funded by a donation from Fondazione Umano Progresso and by donations following the international fundraising initiative by Vittoria Aiello and Massimiliano Albanese (Washington, DC, USA) - donors contributing to this initiative are listed on the site: http://www.massimilianoalbanese.net/ds-research/?lang=en. The fellowship for A.P. has been mainly funded by the Department of Experimental, Diagnostic and Specialty Medicine (DIMES), University of Bologna (Bologna, Italy) and co-funded by the Fondazione Umano Progresso. We are grateful to Kirsten Welter for her expert revision of the manuscript.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

References

1. Borsani G, Ballabio A and Banfi S: A practical guide to orient yourself in the labyrinth of genome databases. Hum Mol Genet 7: 1641-1648, 1998.
2. Pandey A and Lewitter F: Nucleotide sequence databases: A gold mine for biologists. Trends Biochem Sci 24: 276-280, 1999.
3. Baxevanis AD and Bateman A: The importance of biological databases in biological discovery. Curr Protoc Bioinformatics 50: 1.1.1-1.1.8, 2015.
4. Tropp BE (ed): Molecular Biology: Genes to Proteins. 3rd edition. Jones & Bartlett Publishers, Sudbury, MA, 2008.
5. Sambrook J and Russel DW (eds): Molecular Cloning: A Laboratory Manual. Vol 2. 3rd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001.
6. Vitale L, Casadei R, Canaider S, Lenzi L, Strippoli P, D'Addabbo P, Giannone S, Carinci P and Zannotti M: Cysteine and tyrosine-rich 1 (CYYR1), a novel unpredicted gene on human chromosome 21 (21q21.2), encodes a cysteine and tyrosine-rich protein and defines a new family of highly conserved vertebrate-specific genes. Gene 290: 141-151, 2002.
7. Zhang J, Lou X, Shen H, Zellmer L, Sun Y, Liu S, Xu N and Liao DJ: Isoforms of wild type proteins often appear as low molecular weight bands on SDS-PAGE. Biotechnol J 9: 1044-1054, 2014.
8. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al*: Complementary DNA sequencing: Expressed sequence tags and human genome project. Science 252: 1651-1656, 1991.
9. Boguski MS, Lowe TM and Tolstoshev CM: dbEST - database for 'expressed sequence tags'. Nat Genet 4: 332-333, 1993.
10. Nagaraj SH, Gasser RB and Ranganathan S: A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 8: 6-21, 2007.
11. Parkinson J and Blaxter M: Expressed sequence tags: An overview. Methods Mol Biol 533: 1-12, 2009.
12. Gill RW and Sanseau P: Rapid in silico cloning of genes using expressed sequence tags (ESTs). Biotechnol Annu Rev 5: 25-44, 2000.
13. Carulli JP, Artinger M, Swain PM, Root CD, Chee L, Tulig C, Guerin J, Osborne M, Stein G, Lian J, *et al*: High throughput analysis of differential gene expression. J Cell Biochem Suppl 30-31: 286-296, 1998.
14. Sorek R, Shamir R and Ast G: How prevalent is functional alternative splicing in the human genome? Trends Genet 20: 68-71, 2004.
15. Bonizzoni P, Rizzi R and Pesole G: Computational methods for alternative splicing prediction. Brief Funct Genomics Proteomics 5: 46-51, 2006.
16. Brent MR: Genome annotation past, present, and future: How to define an ORF at each locus. Genome Res 15: 1777-1786, 2005.
17. Sanger F: La structure de l'insuline. Bull Soc Chim Biol (Paris) 37: 23-35, 1955 (In French).
18. Yanofsky C, Carlton BC, Guest JR, Helinski DR and Henning U: On the colinearity of gene structure and protein structure. Proc Natl Acad Sci USA 51: 266-272, 1964.
19. Sanger F, Nicklen S and Coulson AR: DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463-5467, 1977.
20. Ruddle FH: The William Allan Memorial Award address: Reverse genetics and beyond. Am J Hum Genet 36: 944-953, 1984.
21. Kozak M: Pushing the limits of the scanning mechanism for initiation of translation. Gene 299: 1-34, 2002.
22. Sambrook J and Russel DW (eds): Rapid amplification of 5' cDNA ends. In: Molecular Cloning: A Laboratory Manual. Vol 3. 3rd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp8.54-8.60, 2001.
23. Okayama H and Berg P: High-efficiency cloning of full-length cDNA. Mol Cell Biol 2: 161-170, 1982.
24. Baralle F: Complete nucleotide sequence of the 5' noncoding region of human alpha-and beta-globin mRNA. Cell 12: 1085-1095, 1977.
25. Proudfoot NJ: Complete 3' noncoding region sequences of rabbit and human beta-globin messenger RNAs. Cell 10: 559-570, 1977.
26. Marotta CA, Wilson JT, Forget BG and Weissman SM: Human beta-globin messenger RNA. III. Nucleotide sequences derived from complementary DNA. J Biol Chem 252: 5040-5053, 1977.
27. Efstratiadis A, Kafatos FC and Maniatis T: The primary structure of rabbit beta-globin mRNA as determined from cloned DNA. Cell 10: 571-585, 1977.
28. Ullrich A, Shine J, Chirgwin J, Pictet R, Tischer E, Rutter WJ and Goodman HM: Rat insulin genes: Construction of plasmids containing the coding sequences. Science 196: 1313-1319, 1977.
29. Casadei R, Strippoli P, D'Addabbo P, Canaider S, Lenzi L, Vitale L, Giannone S, Frabetti F, Facchin F, Carinci P, *et al*: mRNA 5' region sequence incompleteness: A potential source of systematic errors in translation initiation codon assignment in human mRNAs. Gene 321: 185-193, 2003.
30. Harbers M: The current status of cDNA cloning. Genomics 91: 232-242, 2008.
31. Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, *et al*: High-efficiency full-length cDNA cloning by biotinylated CAP trapper. Genomics 37: 327-336, 1996.
32. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, *et al*: CAGE: Cap analysis of gene expression. Nat Methods 3: 211-222, 2006.
33. Frohman MA, Dush MK and Martin GR: Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci USA 85: 8998-9002, 1988.
34. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, *et al*: Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res 17: 746-759, 2007.
35. Suzuki Y, Ishihara D, Sasaki M, Nakagawa H, Hata H, Tsunoda T, Watanabe M, Komatsu T, Ota T, Isogai T, *et al*: Statistical analysis of the 5' untranslated region of human mRNA using 'Oligo-Capped' cDNA libraries. Genomics 64: 286-297, 2000.
36. Porcel BM, Delfour O, Castelli V, De Berardinis V, Friedlander L, Cruaud C, Ureta-Vidal A, Scarpelli C, Wincker P, Schächter V, *et al*: Numerous novel annotations of the human genome sequence supported by a 5'-end-enriched cDNA collection. Genome Res 14: 463-471, 2004.
37. Metzker ML: Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46, 2010.
38. Ingolia NT, Ghaemmaghami S, Newman JR and Weissman JS: Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218-223, 2009.
39. Ingolia NT, Lareau LF and Weissman JS: Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147: 789-802, 2011.
40. Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, Schumann F, Schreiber S, Platzer M, Krawczak M, Hampe J, *et al*: Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res 22: 2208-2218, 2012.
41. Van Damme P, Gawron D, Van Criekinge W and Menschaert G: N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. Mol Cell Proteomics 13: 1245-1261, 2014.
42. Iacono M, Mignone F and Pesole G: uAUG and uORFs in human and rodent 5' untranslated mRNAs. Gene 349: 97-105, 2005.
43. Barbosa C, Peixeiro I and Romão L: Gene expression regulation by upstream open reading frames and human disease. PLoS Genet 9: e1003529, 2013.
44. Nishitani H, Hirose E, Uchimura Y, Nakamura M, Umeda M, Nishii K, Mori N and Nishimoto T: Full-sized RanBPM cDNA encodes a protein possessing a long stretch of proline and glutamine within the N-terminal region, comprising a large protein complex. Gene 272: 25-33, 2001.
45. Kobayashi A, Ito E, Toki T, Kogame K, Takahashi S, Igarashi K, Hayashi N and Yamamoto M: Molecular cloning and functional characterization of a new Cap'n' collar family transcription factor Nrf3. J Biol Chem 274: 6443-6452, 1999.
46. Nomura N, Nagase T, Miyajima N, Sazuka T, Tanaka A, Sato S, Seki N, Kawarabayasi Y, Ishikawa K and Tabata S: Prediction of the coding sequences of unidentified human genes. II. The coding sequences of 40 new genes (KIAA0041-KIAA0080) deduced by analysis of cDNA clones from human cell line KG-1. DNA Res 1: 223-229, 1994.
47. Kingsley C and Winoto A: Cloning of GT box-binding proteins: A novel Sp1 multigene family regulating T-cell receptor gene expression. Mol Cell Biol 12: 4251-4261, 1992.
48. Strippoli P, Pelleri MC, Caracausi M, Vitale L, Piovesan A, Locatelli C, Mimmi MC, Berardi AC, Ricotta D, Radeghieri A, *et al*: An integrated route to identifying new pathogenesis-based therapeutic approaches for trisomy 21 (Down Syndrome) following the thought of Jérôme Lejeune. Sci Postprint 1: e00010, 2013.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

49. Pelleri MC, Cicchini E, Locatelli C, Vitale L, Caracausi M, Piovesan A, Rocca A, Poletti G, Seri M, Strippoli P, *et al*: Systematic reanalysis of partial trisomy 21 cases with or without Down syndrome suggests a small region on 21q22.13 as critical to the phenotype. Hum Mol Genet 25: 2525-2538, 2016.
50. Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, *et al*; Chromosome 21 mapping and sequencing consortium: The DNA sequence of human chromosome 21. Nature 405: 311-319, 2000.
51. Reymond A, Camargo AA, Deutsch S, Stevenson BJ, Parmigiani RB, Ucla C, Bettoni F, Rossier C, Lyle R, Guipponi M, *et al*: Nineteen additional unpredicted transcripts from human chromosome 21. Genomics 79: 824-832, 2002.
52. Pesole G, Gissi C, Grillo G, Licciulli F, Liuni S and Saccone C: Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. Gene 261: 85-91, 2000.
53. Frabetti F, Casadei R, Lenzi L, Canaider S, Vitale L, Facchin F, Carinci P, Zannotti M and Strippoli P: Systematic analysis of mRNA 5' coding sequence incompleteness in *Danio rerio*: An automated EST-based approach. Biol Direct 2: 34, 2007.
54. Casadei R, Piovesan A, Vitale L, Facchin F, Pelleri MC, Canaider S, Bianconi E, Frabetti F and Strippoli P: Genome-scale analysis of human mRNA 5' coding sequences based on expressed sequence tag (EST) database. Genomics 100: 125-130, 2012.
55. Piovesan A, Caracausi M, Pelleri MC, Vitale L, Martini S, Bassani C, Gurioli A, Casadei R, Soldà G and Strippoli P: Improving mRNA 5' coding sequence determination in the mouse genome. Mamm Genome 25: 149-159, 2014.
56. Kochetov AV, Sarai A, Rogozin IB, Shumny VK and Kolchanov NA: The role of alternative translation start sites in the generation of human protein diversity. Mol Genet Genomics 273: 491-496, 2005.
57. Bazykin GA and Kochetov AV: Alternative translation start sites are conserved in eukaryotic genomes. Nucleic Acids Res 39: 567-577, 2011.
58. Ivanov IP, Firth AE, Michel AM, Atkins JF and Baranov PV: Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Res 39: 4220-4234, 2011.
59. Arakaki TL, Pezza JA, Cronin MA, Hopkins CE, Zimmer DB, Tolan DR and Allen KN: Structure of human brain fructose 1,6-(bis)phosphate aldolase: Linking isozyme structure with function. Protein Sci 13: 3077-3084, 2004.
60. Lamour V, Quevillon S, Diriong S, N'Guyen VC, Lipinski M and Mirande M: Evolution of the Glx-tRNA synthetase family: The glutaminyl enzyme as a case of horizontal gene transfer. Proc Natl Acad Sci USA 91: 8670-8674, 1994.
61. Hermann E, Darcissac E, Idziorek T, Capron A and Bahr GM: Recombinant interleukin-16 selectively modulates surface receptor expression and cytokine release in macrophages and dendritic cells. Immunology 97: 241-248, 1999.
62. Schatz G and Dobberstein B: Common principles of protein translocation across membranes. Science 271: 1519-1526, 1996.
63. Nakamura M, Masuda H, Horii J, Kuma K, Yokoyama N, Ohba T, Nishitani H, Miyata T, Tanaka M and Nishimoto T: When overexpressed, a novel centrosomal protein, RanBPM, causes ectopic microtubule nucleation similar to gamma-tubulin. J Cell Biol 143: 1041-1052, 1998.
64. Varshavsky A: The N-end rule: Functions, mysteries, uses. Proc Natl Acad Sci USA 93: 12142-12149, 1996.
65. Rothermel B, Vega RB, Yang J, Wu H, Bassel-Duby R and Williams RS: A protein encoded within the Down syndrome critical region is enriched in striated muscles and inhibits calcineurin signaling. J Biol Chem 275: 8719-8725, 2000.
66. Strippoli P, Petrini M, Lenzi L, Carinci P and Zannotti M: The murine DSCR1-like (Down syndrome candidate region 1) gene family: Conserved synteny with the human orthologous genes. Gene 257: 223-232, 2000.
67. Vitale L, Frabetti F, Huntsman SA, Canaider S, Casadei R, Lenzi L, Facchin F, Carinci P, Zannotti M, Coppola D, *et al*: Sequence, 'subtle' alternative splicing and expression of the CYYR1 (cysteine/tyrosine-rich 1) mRNA in human neuroendocrine tumors. BMC Cancer 7: 66, 2007.
68. Facchin F, Canaider S, Vitale L, Frabetti F, Griffoni C, Lenzi L, Casadei R and Strippoli P: Identification and analysis of human RCAN3 (DSCR1L2) mRNA and protein isoforms. Gene 407: 159-168, 2008.
69. Facchin F, Vitale L, Bianconi E, Piva F, Frabetti F, Strippoli P, Casadei R, Pelleri MC, Piovesan A and Canaider S: Complexity of bidirectional transcription and alternative splicing at human RCAN3 locus. PLoS One 6: e24508, 2011.
70. Casadei R, Pelleri MC, Vitale L, Facchin F, Canaider S, Strippoli P, Vian M, Piovesan A, Bianconi E, Mariani E, *et al*: Characterization of human gene locus CYYR1: A complex multi-transcript system. Mol Biol Rep 41: 6025-6038, 2014.
71. Nagase T, Seki N, Ishikawa K, Tanaka A and Nomura N: Prediction of the coding sequences of unidentified human genes. V. The coding sequences of 40 new genes (KIAA0161-KIAA0200) deduced by analysis of cDNA clones from human cell line KG-1. DNA Res 3: 17-24, 1996.
72. Ribieras S, Lefèbvre O, Tomasetto C and Rio MC: Mouse Trefoil factor genes: Genomic organization, sequences and methylation analyses. Gene 266: 67-75, 2001.
73. Doglio L, Goode DK, Pelleri MC, Pauls S, Frabetti F, Shimeld SM, Vavouri T and Elgar G: Parallel evolution of chordate cis-regulatory code for development. PLoS Genet 9: e1003904, 2013.
74. Hinnebusch AG, Ivanov IP and Sonenberg N: Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science 352: 1413-1416, 2016.
75. Caracausi M, Vitale L, Pelleri MC, Piovesan A, Bruno S and Strippoli P: A quantitative transcriptome reference map of the normal human brain. Neurogenetics 15: 267-287, 2014.
76. Pelleri MC, Piovesan A, Caracausi M, Berardi AC, Vitale L and Strippoli P: Integrated differential transcriptome maps of Acute Megakaryoblastic Leukemia (AMKL) in children with or without Down Syndrome (DS). BMC Med Genomics 7: 63, 2014.
77. Manning AG, Crawford BD, Waskiewicz AJ and Pilgrim DB: unc-119 homolog required for normal development of the zebrafish nervous system. Genesis 40: 223-230, 2004.
78. Piovesan A, Vitale L, Pelleri MC and Strippoli P: Universal tight correlation of codon bias and pool of RNA codons (codonome): The genome is optimized to allow any distribution of gene expression values in the transcriptome from bacteria to humans. Genomics 101: 282-289, 2013.
79. Komar AA: The Yin and Yang of codon usage. Hum Mol Genet 25 (R2): R77-R85, 2016.
80. Piovesan A, Caracausi M, Antonaros F, Pelleri MC and Vitale L: GeneBase 1.1: A tool to summarise data from NCBI gene datasets and its application to an update of human gene statistics. Database (Oxford) 2016: pii: baw153, 2016.
81. Ahsan B, Saito TL, Hashimoto S, Muramatsu K, Tsuda M, Sasaki A, Matsushima K, Aigaki T and Morishita S: MachiBase: A *Drosophila melanogaster* 5'-end mRNA transcription database. Nucleic Acids Res 37 (Database): D49-D53, 2009.
82. Machida RJ and Lin YY: Four methods of preparing mRNA 5' end libraries using the Illumina sequencing platform. PLoS One 9: e101812, 2014.
83. Helbig AO, Gauci S, Raijmakers R, van Breukelen B, Slijper M, Mohammed S and Heck AJ: Profiling of *N*-acetylated protein termini provides in-depth insights into the N-terminal nature of the proteome. Mol Cell Proteomics 9: 928-939, 2010.
84. Doucet A and Overall CM: Amino-Terminal Oriented Mass Spectrometry of Substrates (ATOMS) N-terminal sequencing of proteins and proteolytic cleavage sites by quantitative mass spectrometry. Methods Enzymol 501: 275-293, 2011.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Bioinformatics analysis in genomics

The term "bioinformatics" has only recently been introduced; in fact, it did not appear in literature until 1991, and even then only in the context of the emerging practice of electronic publication. The current concept of "bioinformatics" can probably be best described as the convergence of two technological revolutions: the explosive growth of biotechnology, equalled by that of information technology [Boguski, 1998]. This coincidence is clearly illustrated by the interesting fact that both the size of the DNA GenBank database and the computing power of the computers have doubled at about the same rate (every 18-24 months) for many years. Although the term "bioinformatics" is now very fashionable, many scholars built databases, developed algorithms and made biological discoveries through sequence analysis since the 1970s, long before anyone thought of labelling these activities with a specific term. If any field was referred to, many activities listed today as bioinformatics were included in the context of "molecular evolution".

The specific survey of computer technology in genomics assumes a fundamental value following the advancement of genome projects which aim to determine the complete nucleotide sequence of DNA of various species [Boguski, 1998]. This mass of data provides a privileged raw material for the *ab initio* identification of potential coding sequences as the first step towards the discovery of genes. Another fundamental step was the creation of databases containing partial sequences derived from messenger RNA and obtained from the automated analysis of a large number of bacterial clones obtained from cDNA libraries for many different tissues, in different species. These sequences (EST, expressed sequence tags; Boguski et al., 1993) are available for sequence analysis, which allows you to quickly identify genome sequences that are expressed (genes) without having to deal with the "background" of extragenic sequences, as happens in traditional "genome projects". The availability of these databases, on the one hand, and the evolution of the "software" tools necessary for their analysis, on the other, quickly led to a new approach to gene identification. Computer data became the starting point for *in vitro* experiments ("*in silico*" approach, as computer processors are made of silicon).

The basic operation in this sense consists in comparing the sequences with each other, deducing and quantifying their mutual "similarity" [**Altschul 1998**]. This term, technically referred to as similarity, is purely descriptive of a relationship between the two sequences more significant than that due to chance, while more formally with the term "homology" we mean the notion of a common evolutionary origin of the sequences. In practice, based on the similarity relationships between sequences, it may be possible to infer homology, even if outside of a formal biological model the descent from a common ancestral gene remains hypothetical. Programs that analyze the similarity between sequences are based, in short, on a score assigned based on the number of substitutions, insertions and deletions that must be carried out to convert one sequence into another. The different programs differ in the criteria used in scoring. Currently, the most used program for sequence comparison is based on the BLAST algorithm [Altschul et al., 1997], a heuristic process that identifies similar sequences very quickly, having the specific characteristic of also assigning a value of statistical significance correspondence

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

found. This value ("expect value", or "E" value) corresponds to *the number of comparisons between two sequences with an equal or higher similarity score that could be found, in that particular database, only as a result of chance*; the smaller it is, the more meaningful the match.

For example, by using TBLASTN [**Brenner 1998**] it is possible to start from the amino acid sequence of a known protein, whose coding sequence can be automatically predicted based on the genetic code, and look for similar nucleotide sequences within the EST sequences.
Or it is possible to compare the genomic DNA sequences, determined with the high yield procedures and made publicly available in all the subsequent finishing stages [Ouellette and Boguski, 1997], with the mRNA sequences (known or obtained from the analysis of the EST), for a rapid determination of the genomic structure of genes.
On the other hand, DNA sequences can be compared with amino acid sequences following translation in all the possible protein products they could encode for. Due to the fact that genetic (translation) code is based on three letters, for any given nucleotide sequence there are six possible translation frames (frame +1, +2 and +3 in the query strand and -1, -2 and -3 in the complementary strand). By using the BLASTX ("blast x 6 frames") variant of BLAST one can get clues about the protein coding potential of a nucleotide sequence [**Brenner 1998**].

The availability of databases for many different species also makes it possible to reconstruct the molecular evolution of the sequences of interest, allowing distinguishing between orthology (conservation of a particular gene between different species) and paralogy (presence of a group of homologous genes within a single species). Finally, there are many collections of short sections ("motifs") of amino acid sequences that indicate particular structural or functional elements. Research on these collections from newly identified sequences allows reasonably reliable function predictions to be made [Bork and Gibson, 1996].

Bioinformatics analysis and gene families

DNA sequences in the nuclear diploid genome usually exist in the form of two allelic copies, located on the paternal and maternal homologous chromosomes. In addition to this degree of repetition, about 40% of the human nuclear genome is composed, both in haploid and diploid cells, of groups of closely related non-allelic DNA sequences (families of DNA sequences, or repetitive DNA; Strachan and Read, 1999). Within the considerable variety of repeated DNA sequences, there are also DNA sequence families whose individual members comprise functional genes (multigene families). The operational definition of a family of DNA sequences is the relatively high level of sequence similarity between members of the family, at the level of the whole sequence or its localized regions.

The members of a gene family can be identified by

1. DNA hybridization and cloning, using a gene fragment as a probe for the screening of genetic libraries;

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

2. cloning by amplification with the polymerase chain reaction (PCR), by designing degenerate "primers" that bind to the conserved regions among family members;
3. sequence analysis, which allows the direct calculation of the degree of relationship between the genes.

The fact that two members of a family of DNA sequences show a high degree of similarity is indicative of a common evolutionary origin and is typically related to the conservation of a function.

A large percentage of actively expressed human genes are members of families of DNA sequences; the PFam catalogue [Bateman et al., 2000], maintained at the Sanger Center (Hinxton, Cambridge, UK), classifies 2478 gene families in the version of 2001. We can distinguish different types of gene families. In "classical" gene families, members show a high degree of sequence homology along most of the extent of the genes or, at least, their coding sequence. This characteristic identifies in practice an evolutionary and functional correlation of these sequences. An example is the histone gene families. In some gene families, however, the homology is particularly pronounced within highly conserved regions of the genes, while the similarity between the remaining portions of the coding sequence can be very small. Often these families encode transcription factors that play an important role in the early stages of development, and the conserved sequence encodes a protein domain (folding unit) required for the selective binding of specific target genes to DNA (e.g., the domain Homeobox). Finally, there are also gene families whose members are not obviously correlated at the DNA sequence level, but encode for products characterized by a shared general function and by the presence of conserved short traits ("motifs") of the sequence; for example, the "box" DEAD (amino acid sequence Asp-Glu-Ala-Asp) is found in different genes, whose products all seem to work as RNA helicase. Members of gene families can occasionally be located close to each other in specific subchromosomal regions, such as the genes of the major class I histocompatibility complex (HLA), but are more often found dispersed in the genome.

Many different groups have addressed the problem of grouping protein sequences into families [review in Hofmann, 1998]. The various approaches differ in their degree of automation, in their completeness, in their focus on the complete sequence of proteins or protein domains. Indeed, the relationships between genes and gene families are so complex that "no simple hierarchical scheme can be used to make data easily understandable" [Henikoff et al., 1997], due to the modular composition of proteins.

Among the various tools specifically designed for the reconstruction of gene families through the analysis of the amino acid sequence, of particular importance are PSI-BLAST and programs based on the statistical method Hidden Markov Models (HMM). PSI-BLAST [Altschul et al., 1997] is an "iterative profile-based research". First, a similarity search is performed on a database starting from a single sequence, using BLAST. The significantly similar sequences are aligned to the query sequence, and a "profile" is constructed, a position-specific scoring system derived from the frequency with which a given amino acid residue is observed in a column of the alignment. Since the families of sequences preferentially retain specific residues and critical regions, this information can allow more sensitive research to be carried out, in repeated sequences

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

(iterations). On the other hand, HMM-based programs employ a particular statistical method [for a review see Eddy, 1998a] for the recognition of the configuration of a series of values (the sequence) that can be used to represent the alignment of multiple sequences or sequence segments, to identify the conservation of patterns or individual residues.

The main interest of the study of human gene families consists operationally in obtaining indications on the probable functions of a gene that is similar to a gene already functionally characterized, possibly allowing the recovery of information obtained in model organisms of different species. Despite the progress of large total DNA sequencing projects of different species, many new genes identified to date have not been assigned to gene families. Quoting Hofmann [1998], it can be concluded that "*It might appear that using a combination of domain database searches, BLAST searches and sub-family classification is too much effort for the analysis of a single sequence. However, if one considers how many months of experimental work have been spent on the identification of the protein and the determina- tion of its sequence, it might be worth a few extra hours of computing time too*".

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Fundamentals of database searching

Aligning novel sequences with previously characterized genes or proteins provides important insights into their common attributes and evolutionary origins. The principles underlying the computational tools that can be used to evaluate sequence alignments are discussed.

Stephen F. Altschul
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.
altschul@ncbi.nlm.nih.gov

Efficient DNA sequencing methods make it much easier to obtain information on the amino acid sequence of proteins than on their structures or functions. The sequences of homologous proteins can diverge greatly over time, even though the structure or function of the proteins change little. Thus, much can be inferred about an uncharacterized protein when significant sequence similarity is detected with a well-studied protein. This has been a key motivation for the comparison of DNA and protein sequences. Other goals of sequence comparison include phylogenetic reconstruction and the detection of genes and regulatory regions (see the article by David Haussler on pp. 12–15).

Global and local sequence alignment

Alignments provide a powerful way to compare related sequences, but can be used in an attempt to capture different facts. The alignment of two residues could reflect a common evolutionary origin, or could try to represent common structural roles, which might not always be congruent with evolutionary history. Here, I examine the evolutionary view.

Alignments are generally restricted to describing the most common mutations: insertions, deletions and single-residue substitutions. Insertions or deletions are represented by **null characters**, added to one sequence and aligned with letters in the other; substitutions are represented by the alignment of two different letters. Sequences can be compared by either **global** or **local alignment**, depending on the purpose of the comparison (see Fig. 1). Global alignment forces complete alignment of the input sequences, whereas local alignment aligns only their most similar segments. The method used depends upon whether the sequences are presumed to be related over their entire lengths or to share only isolated regions of homology. Although global and local alignment **algorithms** are reasonably similar, the statistics needed to assess their output are very different.

Alignment scores

To select from the vast number of possible alignments, the standard procedure is to assign them scores; the most frequently used convention is that the higher the score the better the alignment. There are many possible definitions of alignment score, but the most common is simply the sum of scores specified for the aligned pairs of letters, and letters with nulls, of which an alignment consists. A **substitution score** is chosen for each pair of letters that can be aligned; the complete set of these scores is called a **substitution matrix** [PAM (Ref. 1) and BLOSUM (Ref. 2) are the most popular for protein sequence comparisons]. Additionally, scores are chosen for **gaps**, which consist of one or more adjacent nulls in one sequence aligned with letters in the other. Because a single mutational event can insert or delete more than one residue, a long gap should be penalized only slightly more than a short gap. Accordingly, **affine gap costs**, which charge a relatively large penalty for the existence of a gap, and a smaller penalty for each residue it contains, have become the most widely used gap scoring system.

The practical effectiveness of sequence comparison depends critically upon the choice of appropriate substitution and **gap scores**. For **ungapped** local alignments, a complete theory exists describing which substitution scores best distinguish alignments representing true biological relationships from chance similarities. In brief, the score for aligning a given pair of residues i and j depends on the fraction q_{ij} of 'true alignment' positions in which these paired residues tend to appear[3]. Thus, defining a good substitution matrix comes down to estimating the **target frequencies** q_{ij} accurately.

After some thought, it is apparent that the desired target frequencies depend upon the degree of evolution divergence between the related sequences of interest. Thus, what is really required is not a single matrix, but rather a series of matrices tailored to varying degrees of evolutionary divergence[1–3]. This is precisely the perception underlying the construction of the PAM and BLOSUM series of amino acid substitution matrices. These matrices are generally used unmodified for **gapped** local and global alignment. There is no widely accepted theory for selecting gap costs, and their choice has generally been guided by trial and error[4].

0167-7799/98/$ – see front matter. Published by Elsevier Science.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

(a)

```
P00001     1 MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNK---GI   58
               D  KG+ +F     QC T  +  K+  GP L G+ GRK G A G++Y+  N N    G+
P00090     1 Q-DAAKGEAVF----KQCMTCHRADKNMVGPALGGVVGRKAGTAAGFTYSPLNHNSGEAGL   56

P00001    59 IWGEDTLMEYLENPKKYIP--------------GTKMIFVGIKKKEERADLIAYLKKATNE  105
             +W ++ ++ YL +P  Y+                TKM F  +   ++R D+ AYL  AT +
P00090    57 VWTQENIIAYLPDPNAYLKKFLTDKGQADKATGSTKMTF-KLANDQQRKDVAAYL--ATLK  114
```

(b)

```
P13569  1221   EGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRLL-----NTEGEIQIDGVS  1273
               +    ++  +S ++  G+ + L+G +GSGKS   +A L +L      T GEI  DG
P33593    13   QAAQPLVHGVSLTLQRGRVLALVGGSGSGKSLTCAATLGILPAGVRQTAGEILADGKP   70

P13569  1274   WDSITL---------QQWRKAFGVIPQKVFIFSGTFRKNLDPYEQWSDQEIWKVADEV  1322
                 L          Q  R AF  +                +  + + +   + K AD+
P33593    71   VSPCALRGIKIATIMQNPRSAFNPL------------HTMHTHARETCLALGKPADDA  116

P13569  1323   GLRSVIEQFP-GKLDFVLVDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPV  1379
                L + IE         VL      +S G  Q M +A +VL ++  ++ DEP+  LD V
P33593   117   TLTAAIEAVGLENAARVLKLYPFEMSGGMLQRMMIAMAVLCESPFIIADEPTTDLDVV  174
```

Fig. 1. Two protein alignments. **(a)** An optimal global alignment of human cytochrome *c* (105 residues; SWISS-PROT accession number P00001) and *Rhodopseudomonas palustris* cytochrome c_2 (114 residues; SWISS-PROT accession number P00090). **(b)** An optimal local alignment of the human cystic fibrosis transmembrane conductance regulator (1480 residues; SWISS-PROT accession number P13569) and *Escherichia coli* nickel transport ATP-binding protein NIKD (253 residues; SWISS-PROT accession number P33593). Scores for both alignments are calculated using the BLOSUM62 amino acid substitution matrix[2], and affine gap costs[7] that assign the score $-(11 + k)$ to a gap of length k. The global alignment, with score 131, is required to include the whole of the two input sequences and is constructed using the Needleman–Wunsch algorithm[5]. The local alignment, with score 89, involves only those segments of the two input sequences that optimize the score and is constructed using the Smith–Waterman algorithm[6]. On the central line of each alignment, characters indicate identical amino acids and '+' signs indicate similar amino acids (i.e. those whose alignment receives a positive score).

Alignment algorithms and database searches

After defining the score of an arbitrary alignment, one is faced with finding the **optimal** (i.e. highest scoring) alignment, or alignments, of two sequences. Fortunately, given additive scores as discussed above, a set of relatively efficient **dynamic programing** algorithms is available for this task. The first described in the biological literature was the **Needleman–Wunsch algorithm** for global alignment[5]. Subsequently, a slight variant was proposed, termed the **Smith–Waterman algorithm**, which can find the optimal local alignment of two sequences[6]. Both these algorithms require time proportional to the product of the lengths of the sequences being compared. Originally, neither could deal with affine gap costs, but both can now be modified to do so with only a small constant-factor decrease in speed[7].

Because similarities between DNA and protein sequences often span only segments of the sequences involved, the most popular database similarity search programs are based on the Smith–Waterman local alignment algorithm[6]. However, without special-purpose hardware or massively parallel machines the time required by Smith–Waterman renders it too slow for most users. The **FASTA** (http://www2.ebi.ac.uk/fasta3/) (Ref. 8) and **BLAST** (http://www.ncbi.nlm.nih.gov/BLAST) (Refs 9–11) programs therefore use **heuristic** strategies to concentrate their efforts on the sequence regions most likely to be related. Rapid exact-match procedures first identify promising regions, and only then is Smith–Waterman invoked. This approach permits FASTA and BLAST to run 10–100 times faster than full-blown Smith–Waterman, at the cost of overlooking an occasional similarity.

Some of the adjustable parameters of FASTA and BLAST control the details of their heuristics and thus influence the trade-off between speed and sensitivity. The effectiveness of any alignment program depends upon the scoring systems it employs[2–4]. Most importantly, protein similarities corresponding to true homologies are almost always easier to distinguish from chance than their corresponding DNA similarities, so coding DNA should always be conceptually translated to protein before performing a search. The practical use of database search programs is discussed in the article by Steven Brenner on pp. 9–12.

The statistics of alignment scores

To test the biological relevance of a global or local alignment of two sequences, one needs to know how great an alignment score can be expected to occur by chance. In this context, 'chance' can mean the comparison of: (1) real but unrelated sequences; (2) real sequences that are shuffled to preserve compositional properties; or (3) sequences that are generated randomly based upon a DNA or protein sequence model.

Very little of practical value is known about the random distribution of global alignment scores. One of the few ways to evaluate the significance of such a score is to generate an empirical score distribution from the alignment of many 'random' sequences of the same lengths as the two sequences being compared[12]. From this distribution, the ***Z* value** (the number of standard deviations from the mean) for the alignment score of interest can then be estimated. Importantly, it should not be assumed that the score distribution is normal; indeed, its general form is unknown. Therefore, an accurate significance estimate cannot currently be derived from the Z value.

Fortunately, much more is known about the statistics of local alignment scores. Under reasonable assumptions, the random score distribution for optimal ungapped local alignments can be proved to follow an **extreme value distribution**[13,14]. Such a proof is unavailable for gapped local alignments, but computational experiments strongly suggest that the same type of distribution applies[10]. An essential property of the extreme value distribution is that its right-hand tail decays exponentially in x, as opposed to x^2 for the normal distribution. Improperly assuming a normal distribution for optimal local alignment scores can thus result in gross exaggerations of statistical significance.

Current versions of the FASTA and BLAST search programs report the **raw scores** of the alignments they return, as well as assessments of their statistical significance, based upon the extreme value distribution. Most simply, these assessments take the form of ***E* values**. The E value for a given alignment depends upon its score, as well as the lengths of both the query sequence and the database searched. It represents the number of distinct alignments with equivalent or superior score that might have been expected to have occurred purely by chance. Thus an E value of five is not statistically significant, whereas an E value of 0.01 is. BLAST also reports **bit scores**, which are scaled versions of the raw scores[11]. A bit score takes into account the statistical

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

parameters[3,10,13] of the scoring system employed, and is therefore more informative than a raw score for describing the quality of an alignment.

Masking regions of restricted composition

Many DNA and protein sequences contain regions of highly restricted nucleic acid and amino acid composition and regions of short elements repeated many times[15]. The standard alignment models and scoring systems were not designed to capture the evolutionary processes that led to these **low-complexity regions**. As a result, two sequences containing compositionally biased regions can receive a very high similarity score that reflects this bias alone. For many purposes, these regions are uninteresting and can obscure other important similarities. Therefore, programs that filter low-complexity regions from query or database sequences will often turn a useless database search into a valuable one[15].

Multiple sequences

Global and local pairwise sequence comparison and alignment can be generalized to multiple sequences. From multiple alignments, **profiles** [related to hidden Markov models (**HMMs**)] can be abstracted and these can greatly enhance the sensitivity of database search methods to evolutionarily distant and subtle sequence relationships[11]. As discussed by Sean Eddy on pp. 15–18 and by Kay Hofmann on pp. 18–21, this area is increasingly becoming the focus of algorithm and database development for biological sequence comparison.

Dedication

This article is dedicated to Dr Bruce W. Erickson, friend and mentor.

References

1 Dayhoff, M.O. *et al.* (1978) in *Atlas of Protein Sequence and Structure* (Vol. 5, Suppl. 3) (Dayhoff, M.O., ed.), pp. 345–352, National Biomedical Research Foundation
2 Henikoff, S. and Henikoff, J.G. (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 10915–10919
3 Altschul, S.F. (1991) *J. Mol. Biol.* 219, 555–565
4 Pearson, W.R. (1995) *Protein Sci.* 4, 1145–1160
5 Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Biol.* 48, 443–453
6 Smith, T.F. and Waterman, M.S. (1981) *J. Mol. Biol.* 147, 195–197
7 Gotoh, O. (1982) *J. Mol. Biol.* 162, 705–708
8 Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448
9 Altschul, S.F. *et al.* (1990) *J. Mol. Biol.* 215, 403–410
10 Altschul, S.F. and Gish, W. (1996) *Methods Enzymol.* 266, 460–480
11 Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
12 Fitch, W.M. (1983) *J. Mol. Biol.* 163, 171–176
13 Karlin, S. and Altschul, S.F. (1990) *Proc. Natl. Acad. Sci. U. S. A.* 87, 2264–2268
14 Dembo, A. *et al.* (1994) *Ann. Probab.* 22, 2022–2039
15 Wootton, J.C. and Federhen, S. (1993) *Comput. Chem.* 17, 149–163

Practical database searching

Sequence comparisons need to be performed as carefully as wet-lab procedures, in terms of both experimental design and interpretation. The basic requirements of database searching, the factors that can affect the search results and, finally, how to interpret the results are discussed.

More sequences have been putatively characterized by database searches than by any other single technology. For good reason: programs like **BLAST** are fast and reliable. However, sequence comparison procedures should be treated as experiments analogous to standard laboratory procedures. Their use deserves the same care both in the design of the experiment and in the interpretation of results.

Steven E. Brenner
Dept of Structural Biology, Stanford University, Stanford, CA 94305-5400, USA.
brenner@hyper.stanford.edu

The database search experiment

Design of a BLAST database search requires consideration of what information is to be gained about the query sequence of interest. The main constraint is that database searching can only reveal similarity. However, from this similarity, homology (i.e. evolutionary relationship) can be inferred and, from that, one might be able to infer function. Although the former inference is now reliable

0167-7799/98/$ – see front matter © 1998 Elsevier Science. All rights reserved.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

parameters[3,10,13] of the scoring system employed, and is therefore more informative than a raw score for describing the quality of an alignment.

Masking regions of restricted composition

Many DNA and protein sequences contain regions of highly restricted nucleic acid and amino acid composition and regions of short elements repeated many times[15]. The standard alignment models and scoring systems were not designed to capture the evolutionary processes that led to these **low-complexity regions**. As a result, two sequences containing compositionally biased regions can receive a very high similarity score that reflects this bias alone. For many purposes, these regions are uninteresting and can obscure other important similarities. Therefore, programs that filter low-complexity regions from query or database sequences will often turn a useless database search into a valuable one[15].

Multiple sequences

Global and local pairwise sequence comparison and alignment can be generalized to multiple sequences. From multiple alignments, **profiles** [related to hidden Markov models (**HMMs**)] can be abstracted and these can greatly enhance the sensitivity of database search methods to evolutionarily distant and subtle sequence relationships[11]. As discussed by Sean Eddy on pp. 15–18 and by Kay Hofmann on pp. 18–21, this area is increasingly becoming the focus of algorithm and database development for biological sequence comparison.

Dedication

This article is dedicated to Dr Bruce W. Erickson, friend and mentor.

References

1 Dayhoff, M.O. *et al.* (1978) in *Atlas of Protein Sequence and Structure* (Vol. 5, Suppl. 3) (Dayhoff, M.O., ed.), pp. 345–352, National Biomedical Research Foundation
2 Henikoff, S. and Henikoff, J.G. (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 10915–10919
3 Altschul, S.F. (1991) *J. Mol. Biol.* 219, 555–565
4 Pearson, W.R. (1995) *Protein Sci.* 4, 1145–1160
5 Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Biol.* 48, 443–453
6 Smith, T.F. and Waterman, M.S. (1981) *J. Mol. Biol.* 147, 195–197
7 Gotoh, O. (1982) *J. Mol. Biol.* 162, 705–708
8 Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448
9 Altschul, S.F. *et al.* (1990) *J. Mol. Biol.* 215, 403–410
10 Altschul, S.F. and Gish, W. (1996) *Methods Enzymol.* 266, 460–480
11 Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
12 Fitch, W.M. (1983) *J. Mol. Biol.* 163, 171–176
13 Karlin, S. and Altschul, S.F. (1990) *Proc. Natl. Acad. Sci. U. S. A.* 87, 2264–2268
14 Dembo, A. *et al.* (1994) *Ann. Probab.* 22, 2022–2039
15 Wootton, J.C. and Federhen, S. (1993) *Comput. Chem.* 17, 149–163

Practical database searching

Sequence comparisons need to be performed as carefully as wet-lab procedures, in terms of both experimental design and interpretation. The basic requirements of database searching, the factors that can affect the search results and, finally, how to interpret the results are discussed.

More sequences have been putatively characterized by database searches than by any other single technology. For good reason: programs like **BLAST** are fast and reliable. However, sequence comparison procedures should be treated as experiments analogous to standard laboratory procedures. Their use deserves the same care both in the design of the experiment and in the interpretation of results.

Steven E. Brenner
Dept of Structural Biology, Stanford University, Stanford, CA 94305-5400, USA.
brenner@hyper.stanford.edu

The database search experiment

Design of a BLAST database search requires consideration of what information is to be gained about the query sequence of interest. The main constraint is that database searching can only reveal similarity. However, from this similarity, homology (i.e. evolutionary relationship) can be inferred and, from that, one might be able to infer function. Although the former inference is now reliable

0167-7799/98/$ – see front matter © 1998 Elsevier Science. All rights reserved.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Box 1. Database searching: basic considerations

- Think about every step
- Search a large current database
- Compare as protein rather than DNA
- Filter query for low-complexity regions
- Interpret scores with *E* values
- Recognize that most homologs are not found by pairwise sequence comparison
- Consider slower and more powerful methods, but use iterative programs with great caution

for carefully performed sequence comparison, the second is still fraught with challenges. Box 1 provides some guidelines for performing reliable and sensitive database searches.

Planning a good experiment requires an understanding of the method being applied. Fundamentally, database searches are a simple operation: a query sequence is locally aligned with each of the sequences (called targets) in a database. Most programs, such as BLAST (Ref. 1) and **FASTA** (Ref. 2), use **heuristics** to speed up the alignment procedure, while the **Smith–Waterman algorithm**[3] (implemented, for example, in **SSEARCH**) rigorously compares the query sequence with each target in the database.

A score is computed from each alignment, and the query–target pairs with the best scores are then reported to the user. Typically, statistics are used to help improve the interpretation of these scores. A more detailed description of the process can be found in the article by Stephen Altschul on pp. 7–9. Although BLAST is the most widely used tool for sequence comparison, many other programs can help identify, confirm and interpret distant evolutionary relationships.

Databases, programs and comparison types

Formulation of the experiment begins with a decision about what types of sequences to compare: DNA, protein or DNA as protein. If the sequence under consideration is a protein or codes for a protein, then the search should probably take place at the protein level, because proteins allow one to detect far more distant homology than does DNA[2,4]. For example, in DNA comparisons, there is noise from the rapidly mutated third-base position in each codon and from comparisons of noncoding frames (although this latter issue still arises in DNA-as-protein searches). In addition, amino acids have chemical characteristics that allow degrees of similarity to be assessed rather than simple recognition of identity or non-identity. For these reasons, DNA versus DNA comparison (using the blastn program) is typically only used to find identical regions of sequence in a database. One would carry out such a search to discover whether the gene has been previously sequenced and to determine where it is expressed or where splice junctions occur. In short, protein-level searches are valuable for detecting evolutionarily related genes, while DNA searches are best for locating nearly identical regions of sequence.

Next, it is necessary to select a database to search against. For homology searches, the most commonly searched database on the NCBI (National Center for Biotechnology Information) website is the **nr database**. The nr protein database combines data from several sources, removes the redundant identical sequences and yields a collection with nearly all known proteins. The NCBI nr database is frequently updated in order to incorporate as many sequences as possible. Obviously, a search will not identify a sequence that has not been included in the database and, as databases are growing so rapidly, it is essential to use a current database. Several specialized databases are also available, each of which is a subset of the nr database. ***E*-value** statistics (discussed below) are affected by database size, so, if you are interested in searching for proteins of known structure, it is best to just search the smaller **pdb database**.

One might also wish to search DNA databases at the protein level. Programs can do so automatically by first translating the DNA in all six reading frames and then making comparisons with each of these conceptual translations. The nr DNA database, which contains most known DNA sequences except **GSS**s, **EST**s, **STS**s and **HTGS**s, is useful to search when hunting new genes; the identified genes in this database would already be in the protein nr database. Searches against the GSS, EST, STS and HTGS databases can find new homologous genes and are especially useful for learning about expression data or genome map location.

Because of the different combinations of queries and database types, there are several variants of BLAST (see Table 1). Note that it is desirable to use the newest versions of BLAST, which support **gapped alignments** (see the article by Stephen Altschul on pp. 7–9). The older versions are slower, detect fewer homologs and have problems with some statistics. The programs can be run over the World Wide Web (WWW) and can be downloaded from an **ftp** site to run locally. Another option is to use the FASTA package[2]. The FASTA program can be slower but more effective than BLAST. The package also contains SSEARCH, an implementation of the rigorous Smith–Waterman algorithm, which is slow but the most sensitive. As described in the article by Sean Eddy on pp. 15–18, iterative programs such as **PSI-BLAST** require extreme care in their operation because they can provide very misleading results; however, they have the potential to find more homologs than purely pairwise methods.

Filtering

The statistics for database searches assume that unrelated sequences will look essentially random with respect to each other. However, certain patterns in sequences violate this rule. The most common exceptions are long runs of a small number of different residues (such as a poly-alanine tract). Such regions of sequence could spuriously obtain extremely high match scores. For this reason, the NCBI BLAST server will automatically remove such sections in proteins (replacing them with an X) using the **SEG** program[5] if 'default **filtering**' is selected. DNA sequences will be similarly **masked** by **DUST**.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Table 1. BLAST variants for different searches[a]

Program	Query	Database	Comparison	Common use
blastn	DNA	DNA	DNA level	Seek identical DNA sequences and splicing patterns
blastp	Protein	Protein	Protein level	Find homologous proteins
blastx	DNA	Protein	Protein level	Analyze new DNA to find genes and seek homologous proteins
tblastn	Protein	DNA	Protein level	Search for genes in unannotated DNA
tblastx	DNA	DNA	Protein level	Discover gene structure

[a]Similar variant programs are available for FASTA. Protein-level searches of DNA sequences are performed by comparing translations of all six reading frames.

Although these programs automatically remove the majority of problematic matches, some problems invariably slip through; moreover, valid hits might be missed if part of the sequence is masked. Therefore, it might be helpful to try using different masking parameters.

Other sorts of filtering are also often desirable. For example, **iterative searches** are prone to contamination by regions of proteins that resemble coiled coils or transmembrane helices. The problem is that a protein that is similar only in these general characteristics might match initially. The profile then emphasizes these inappropriate characteristics, eventually causing many spurious hits. Heavily cysteine-rich proteins can also obtain anomalous high scores. Especially if these characteristics are not filtered, it is necessary to review the alignment results carefully to ensure that they have not led to incorrect matches.

Alignment, algorithmic and output parameters

Three other sets of parameters also affect search results, but they rarely require careful consideration by most users. First, the matrix and gap parameters determine how similarity between two sequences is determined. When two residues in a protein are aligned, programs use the matrix to determine whether the amino acids are similar (and thus receive a positive score) or very different. The default matrix for BLAST is called BLOSUM62 (Ref. 6), and the programs will not currently operate reliably with other matrices. The gap parameters determine how much an alignment is penalized for having gaps: the existence parameter is a fixed cost for having a gap and the per-position cost is a cost dependent upon the length (i.e. the number of residues). Typically, there is a large cost associated with introducing a gap and a small additional cost such that longer gaps are worse. It is rarely very beneficial to change these from their defaults.

The second set of parameters determines the heuristics that BLAST uses. By altering these numbers, it is possible to make the program run slower and be more sensitive, or to run faster at the cost of missing more homologs. The complexity of these parameters in BLAST precludes extensive description here. Currently, it is very rare for users to alter these options from the defaults. The FASTA program has one such parameter, called **ktup**, that a user will often want to set. Searches with ktup = 1 are slower, but more sensitive, than BLAST; ktup = 2 is fast, but less effective.

A third set of parameters regulates how many results are reported. By default, the programs will report only matches with an ***E* value** (described below) up to 10. The total number of matches is limited to the best 500, and detailed information with the alignment is provided for up to 100 pairs. To retrieve more matches, these numbers can be increased.

Interpretation of results

Interpretation of the results of a sequence database search involves first evaluating the matches to determine whether they are significant and therefore imply homology. The most effective way of doing this is through use of statistical scores or *E* values. The *E* values are more useful than the **raw** or **bit scores**, and they are far more powerful than percentage identity (which is best not even considered unless the identity is very high)[7]. Fortunately, the *E* values from FASTA, SSEARCH and NCBI gapped BLAST seem to be accurate and are therefore easy to interpret (see Ref. 7).

The *E* value of a match should measure the expected number of sequences in the database that would achieve a given score by chance. Therefore, in the average database search, one expects to find ten random matches with *E* values below 10; obviously, such matches are not significant. However, lacking better matches, sequences with these scores might provide hints of function or suggest new experiments. Scores below 0.01 would occur by chance only very rarely and are therefore likely to indicate homology, unless biased in some way. Scores of near 1e−50 (1×10^{-50}) are now seen frequently, and these offer extremely high confidence that the query protein is evolutionarily related to the matched target in the database.

Inferring function from the homologous matched sequences is a problematic process. If the score is extremely good and the alignment covers the whole of both proteins, then there is a good chance that they will share the same or a related function. However, it is dangerous to place too much trust in the query having the same function as the matched protein; functions do diverge, and organismal or cellular roles can alter even when biochemical function is unchanged. Moreover, a significant fraction of functional annotations in databases are wrong, so one needs to be careful. There are other complexities; for example, if only a portion of the proteins align, they might share a domain that only contributes one aspect of the overall function. It is often the case that all of the highest-scoring hits align to one region of the query, and matches to other regions need to be sought much lower

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

URLS

BLAST Web site
http://www.ncbi.nlm.nih.gov/BLAST/

BLAST FTP site
ftp://ncbi.nlm.nih.gov/blast/

FASTA at EBI
http://www2.ebi.ac.uk/fasta3/

FASTA FTP site
ftp://ftp.virginia.edu/pub/fasta

Sequence search site
http://sss.stanford.edu/sss/

in the score ranking. For this reason, it is necessary to consider carefully the overlap between the query and each of the targets.

Database search methods are also limited because most homologous sequences have diverged too far to be detected by pairwise sequence comparison methods[7]. Thus, failure to find a significant match does not indicate that no homologs exist in the database; rather, it suggests that either more-powerful computational methods (such as those described by Sean Eddy on pp. 15–18 and by Kay Hofmann on pp. 18–21) or experiments would be necessary to locate them.

Conclusion

One should neither have excessive faith in the results of a BLAST run nor blithely disregard them. The BLAST programs are well-tested and reliable indicators of sequence similarity, and their underlying principles are straightforward. Complexities added by the fast algorithms typically need not be carefully considered, because the program and its parameters have been optimized for hundreds of thousands of daily runs. If one is careful about posing the database search experiment and interprets the results with care, sequence comparison methods can be trusted to provide an incomparable wealth of biological information rapidly and easily.

References

1 Altshul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
2 Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448
3 Smith, T.F. and Waterman, M.S. (1981) *J. Mol. Biol.* 147, 195–197
4 States, D.J. *et al.* (1991) *Methods* 3, 66–70
5 Wootton, J.C. (1994) *Comput. Chem.* 18, 269–285
6 Henikoff, S. and Henikoff, J.G. (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 10915–10919
7 Brenner, S.E., Chothia, C. and Hubbard, T.J.P. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 6073–6078

Computational genefinding

A major challenge in the analysis of genomic DNA sequence is to find the functional sites that encode elements responsible for gene structure, regulation and transcription. A variety of computational tools can help to isolate the 'signal' from the 'noise'.

David Haussler
Computer Science Dept,
University of California, Santa Cruz,
CA 95064, USA.
haussler@cse.ucsc.edu

Expanded version of this article
http://www.cse.ucsc.edu/~haussler/pubs.html

Computational methodology for finding genes and other functional sites in genomic DNA has evolved significantly over the past 20 years (for reviews, see Refs 1–3). The genomic elements that researchers seek include splice sites, start and stop codons, branch points, promoters and terminators of transcription, polyadenylation sites, ribosomal binding sites, topoisomerase II binding sites, topoisomerase I cleavage sites and various transcription factor binding sites[4]. Local sites such as these are called **signals**, and methods for detecting them can be called signal sensors. In contrast, extended and variable-length regions, such as exons and introns, are called **contents** and are recognized by methods that can be called content sensors[5].

Signal sensors

The most basic signal sensor is a simple consensus sequence, or an expression that describes a consensus sequence together with allowable variations. More sensitive sensors can be designed using **weight matrices** in place of the consensus, in

0167-7799/98/$ – see front matter © 1998 Elsevier Science. All rights reserved.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

CHAPTER ONE

Genetics and genomics of Down syndrome

Pierluigi Strippoli, Maria Chiara Pelleri, Allison Piovesan*, Maria Caracausi, Francesca Antonaros and Lorenza Vitale

Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, Bologna, Italy
*Corresponding author: E-mail: allison.piovesan2@unibo.it

Contents

Abstract

Down Syndrome (DS) is the most frequent form of intellectual disability (ID) of genetic origin, whose main features include craniofacial dysmorphisms and cardiovascular defects. In 1959, Lejeune and coll. described an extra copy of chromosome 21 (Hsa21) in children with DS (trisomy 21, or T21). We first review how different biological mechanisms may lead to the gain of genetic material of Hsa21 in the cells, originating from different combinations of genetic conditions, including a free or translocated extra copy of Hsa21, distributed in all cells or only in a part of them (mosaicism), with a complete or partial representation of the Hsa21 long arm (21q). Although it is broadly

International Review of Research in Developmental Disabilities, Volume 56
ISSN 2211-6095
https://doi.org/10.1016/bs.irrdd.2019.06.001

© 2019 Elsevier Inc.
All rights reserved.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

agreed that the DS phenotype originates from the altered expression of the genes located on Hsa21, its molecular pathogenesis is still unknown. We therefore illustrate how recent genomic science may be useful in the elucidation of the genotype-phenotype relationship in DS.

1. Introduction

Down syndrome (DS) is the most common genetic cause of intellectual disability (ID) of genetic origin in humans, being observed in 1 in ~700 live births (Parker et al., 2010), while a spontaneous fetal loss rate of 43% based on data between the time of chorionic villus sampling and term (Morris, Wald, & Watt, 1999) allows the estimation of 1 in ~400 conceptions.

While the most constant and typical features are ID and craniofacial dysmorphisms, there is the possibility of an increased frequency of many other signs and symptoms in a variety of organs and systems (Hickey, Hickey, & Summar, 2012). In this review we will focus on the genetic bases of the syndrome and on the recent developments of genomic science in understanding the genotype-phenotype relationship in DS. In particular, the identification of a critical region for the main features of DS will be discussed in relationship to a better understanding of the mechanisms leading to ID.

2. History

It is likely that some artistic representations depict individuals with physical characteristics of DS in the Greco-Roman world, in many Central and South American pre-Columbian cultures, in Cambodia's Khmer temples, as well as in European Renaissance paintings (Levitas & Reid, 2003; Martinez-Frias, 2005; Stahl & Tourame, 2013).

Jean-Etienne-Dominique Esquirol (1838) and Édouard Séguin (1846) were probably the first physicians to describe the clinical picture that was later associated with trisomy 21 (Esquirol, 1838; Séguin, 1846; reviewed in Roubertoux & Kerdelhue, 2006). Esquirol described subjects with oblique eye fissures, epicanthic eye-folds, a flat nasal bridge, a protruding tongue, and short stature, along with ID. Séguin named the condition "furfuraceous idiocy".

In his 1866 description of children with DS, British doctor John Langdon Down, for whom the disorder was later named, correctly focused on the

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

most constant features, i.e. ID and a typical facies, although summarized them as "idiocy of Mongolian type", which reflected conceptions of his time but is unacceptable today (Down, 1866). Down used an obsolete racial framework to classify children with ID, assuming a regression when showing Mongolian traits (Roubertoux & Kerdelhue, 2006). Actually, before the development of modern genetic analysis, there was no means to demonstrate an etiology for the syndrome. Therefore, even in his 1945 review about "the causation of mongolism", Engler listed uterine exhaustion, alcoholism, syphilis, increased amniotic pressure and ovum nesting into damaged mucous membrane of the uterus (Engler, 1945) as possible causes, all ruled out later.

This situation makes the historical relevance of the first description of a genetic anomaly in humans clear. In January 1959, the young French doctor Jérôme Lejeune (1926–94) and coll. (Lejeune, Gauthier, & Turpin, 1959a) demonstrated that "mongolism" was associated with the presence of an additional copy of human chromosome 21 (Hsa21) in all three children they had studied. Two months later a similar report for nine children was published (Lejeune, Gautier, & Turpin, 1959b), followed by the 1961 proposal published by the journal "The Lancet" to abandon the term "mongolism" and to use "Down syndrome" or "trisomy 21" (T21) to refer to this condition (Allen et al., 1961; Stevenson, 2009). In his 1954 review "Observations on the etiology of mongolism" published in "The Lancet" (Penrose, 1954), Penrose cited two works coauthored by Reymond Turpin and Lejeune in 1953 and 1954 as clues pointing to the relevance of a genetic background for DS, in particular due to the dermatoglyphes that are characteristic and known to be determined at early stages of development, thus providing an intellectual framework for the study of chromosomes in this condition as soon as such a study became technically feasible.

The discovery of this genetic mutation leading to the presence of three copies of Hsa21, instead of the normal two, in the cells of the affected individuals is commonly recognized as a milestone in the history of genetics (National Human Genome Research Institute, 2016). For the first time, a given clinical picture is connected to a specific alteration of the human DNA, providing the basis of the field of medical genetics. Remarkably, the discovery of T21 had relevant social consequences for affected children whose parents were no longer suspected to be alcoholics or infected with syphilis, while also paving the way for the scientific research on the condition.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

3. Genetic causes and types of trisomy 21

Given the prevalence of DS discussed above, T21 can be considered the most common genetic anomaly in humans. It is typically a *de novo* mutation presenting without predictability and independently of environmental or genetic factors in the parents. While different risk factors for DS including folate metabolism (Coppedè, 2015) as well as dietary, lifestyle, environmental, occupational, genetic, and epigenetic factors (reviewed by Cocchi et al., 2010; Coppedè, 2016; Morris, De Vigan, Mutton, & Alberman, 2005) have been suggested over time, there is no clear model explaining the birth of a child with DS.

The only certain factor increasing the probability of having a child with DS to date is maternal age (Cocchi et al., 2010; Morris et al., 2005). This was actually recognized well before the discovery of T21, and the association of DS with advanced maternal age was described in 1876 by Fraser and Mitchell according to Jarvik and coll. (Fraser & Mitchell, 1876; Jarvik, Falek, & Pierson, 1964) and in 1895 by Shuttleworth as noted by Lejeune (Lejeune, 1963; Shuttleworth, 1900). It was also supported by this statement written just a few years before the discovery of T21: "The only aetiological fact concerning mongolism which has hitherto stood out clearly and consistently in all carefully conducted inquiries is the relationship to maternal age" (Penrose, 1954).

In a recent revised estimation on the subject, odds have been observed to increase from 1:959 at the age of 30 to 1:84 at the age of 40 (Morris et al., 2005), although a high proportion of children with DS is born to young mothers who have a proportionally higher delivery rate. It may be considered that female newborn ovaries have about 700.000 oocytes at birth and following a continuous decline in this number and absence of oogenesis, a typical ovarian reserve at puberty includes about 300.000 oocytes (Hartshorne, Lyrakou, Hamoda, Oloto, & Ghafari, 2009). Therefore, a longer permanence in the ovary before ovulation may be associated with a higher probability of accumulating biological damage over time (Kurahashi et al., 2012). It has been proposed that there is an accumulation of T21 cells at prenatal and postnatal oocyte development (germinal T21 mosaicism), possibly leading to a higher proportion at later maternal ages (Hulten, Oijerstedt, Iwarsson, & Jonasson, 2014), although the molecular mechanisms involved in maternal Hsa21 malsegregation during oogenesis are still not fully understood (Coppedè, 2015). It has been alternatively proposed

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

that the age effect occurs simply because the older ovary is less efficient at segregating susceptible bivalents (paired homologous chromosomes) established in the fetal ovary (Hassold & Hunt, 2001).

A recent study of live human oocytes identified an error-prone spindle assembly mechanism mediated by chromosomes, rather than by centrosome or other microtubule organizing centers, as a major contributor to chromosome segregation defects typical of human oocytes (Holubcova, Blayney, Elder, & Schuh, 2015). Moreover, in meiosis I, sister chromatids must attach to spindle kinetochore fibers emanating from the same pole for a successful separation of chromosome homologs, but while in mouse and budding yeast sister kinetochores remain closely associated and act as a unified structure, in humans individual kinetochores within a pair can form independent attachments to spindle kinetochore fibers, making human oocytes more prone to aneuploidy while with increasing female age the separation between kinetochores further increases (Patel, Tan, Hartshorne, & McAinsh, 2015). As recognized by Lejeune and coll. in their 1959 article, human meiosis is instead in this regard similar to that of the fruit fly, *Drosophila melanogaster*: "As we know in Drosophila, non-disjunction is strongly affected by maternal aging, so that such a mechanism would reflect the increased frequency of Mongolism according to the advanced age of the mother" (Lejeune, Gautier, et al., 1959a, our translation).

Although an increase in older mothers is observed, the prevalence of DS births remains stable in most registers as a result of the increasing use of prenatal diagnostic procedures followed by elective termination of pregnancy (Cocchi et al., 2010). A recent study found 1 in 806 live births of children with DS in Massachusetts in comparison to the estimated frequency of 1 in 408 if DS-related elective pregnancy terminations were not considered (de Graaf, Buckley, & Skotko, 2016).

Whatever the original cause of the genetic error might be, an error in the distribution or copy of Hsa21 molecules results in an extra copy of Hsa21 in the cells of affected people. Four main types of errors may be related to T21, depending on whether the error occurs in meiosis or mitosis or involves a translocation or DNA replication.

1. Meiosis. Following their first descriptions of T21, Lejeune and coll. readily hypothesized that the likely cause was an error in meiosis, causing an imbalance in the distribution of the chromosomes (Lejeune, Gautier, Turpin, 1959b; Lejeune, Turpin, & Gautier, 1959b) (Fig. 1). They wrote, naming Hsa21 as "Vh" following an alternative nomenclature (Lejeune, Turpin, & Gautier, 1959a; Robinson, 1960): "The origin of this trisomy

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

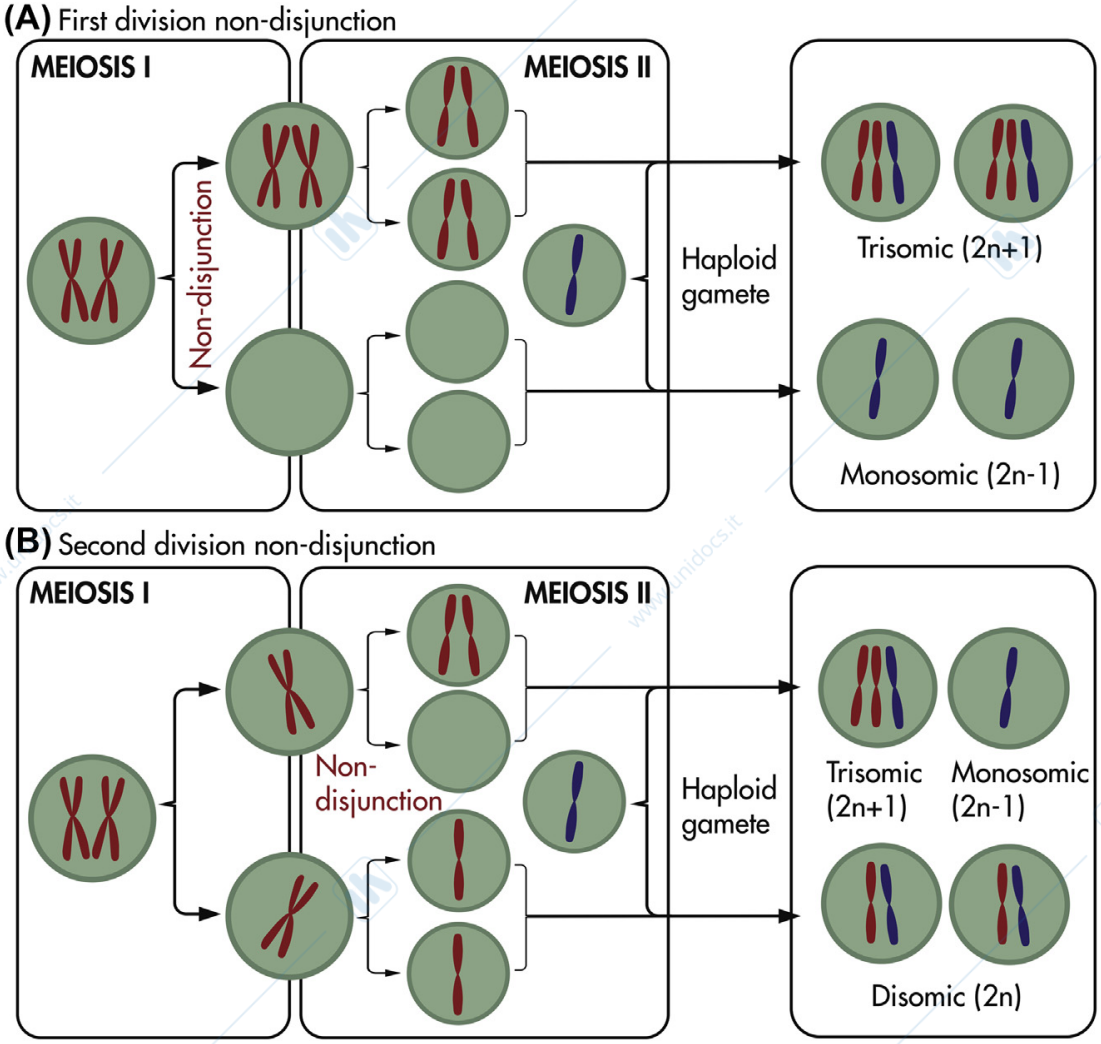


Fig. 1 Chromosome non-disjunction in meiosis (only the involved pair is represented). *Image courtesy of Marcello Fedi.*

could be explained by the non-disjunction of two Vh chromosomes joined at the beginning of meiosis. Thereby, the resulting egg would be found once on two diploids for the chromosome Vh though normally haploid for the rest of the genome. Consequently, fertilization by a normal sperm would lead to trisomy for the chromosome Vh" (Lejeune, Turpin, et al., 1959b; our translation from French). At the time, a demonstration of the expected counterpart of such a mechanism (newborns with monosomy) was lacking. The subsequent demonstration that fetuses with autosomal monosomy exist but are spontaneously aborted (Kelly, Almy, Jakovic, & Buckner, 1965) being non-viable, allowed full confirmation of the non-disjunction hypothesis.

In theory, failure in the separation of chromosomes causing T21 could occur in paternal or maternal meiosis, during anaphase of meiosis I (non-disjunction of bichromatid homologous chromosomes, i.e. failure in the

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

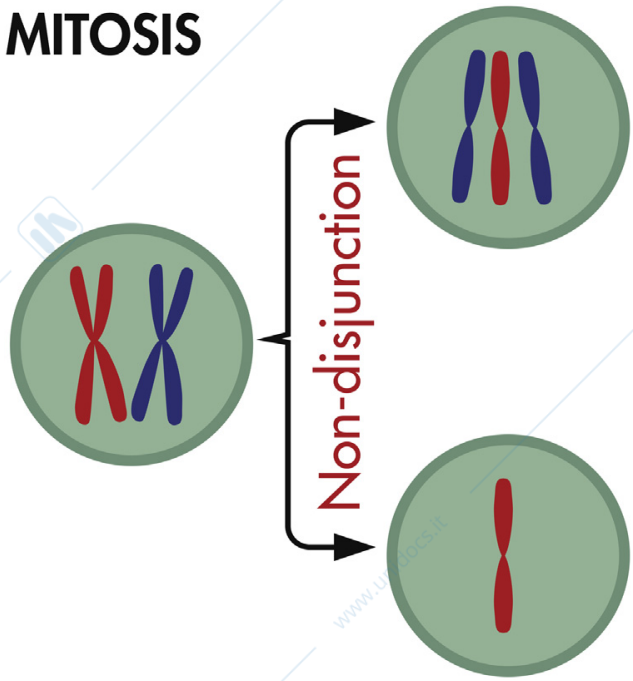


Fig. 2 Chromosome non-disjunction in mitosis (only the involved pair is represented). *Image courtesy of Marcello Fedi.*

separation of the two homologs of each chromosome pair so that they are not allowed to move toward opposite poles of the cell) or, alternatively, during anaphase of meiosis II (non-disjunction of sister chromatids of one bichromatid chromosome) (Strachan & Read, 2019) (Fig. 1).

2. Mitosis. On the other hand, if the error in the distribution of the members of the Hsa21 pair occurs during anaphase of mitosis of a somatic cell of the embryo, one monosomic and one trisomic daughter cell will be produced (Fig. 2). While monosomic cells can't survive, the trisomic cell line will proliferate together with the normal cells so that the subject will be a "mosaic" of cells with different karyotypes, carrying a part of cells with T21 and a part euploid. The timing of the postzygotic mitotic non-disjunction during development will determine the percentage of trisomic cells (the earlier the error, the higher the proportion of aneuploid cells).

The possibility of errors in mitosis of a developing embryo also explains the case that two monozygotic twins might be found discordant for DS (Van Beukering & Vervoorn, 1956) because they are discordant for the aneuploidy (Lejeune & Aurias, 1976) if the mutation occurs following the separation of the two embryos. Several Authors have reported monozygotic twins, one with typical development and the other with T21 (Chang et al., 2017; Choi et al., 2013; Citterio, 1964; Dahoun et al., 2008; Dekaban, 1965; Lejeune et al., 1962; Macatangga, De la Calle, Torres, & Bartha,

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

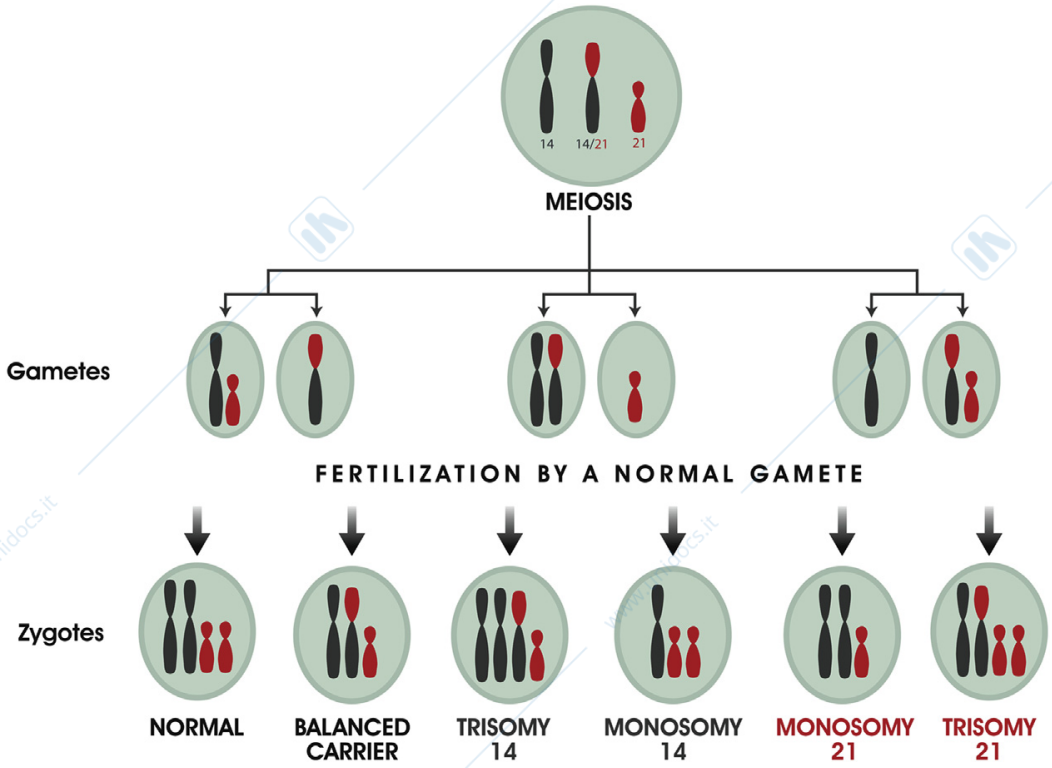


Fig. 3 Robertsonian translocation. *Image courtesy of Marcello Fedi.*

2016; O'Donnell, Pertile, Sheffield, & Sampson, 2004; Rogers, Voullaire, & Gold, 1982).

3. Translocation. The fusion of the long arms of two acrocentric chromosomes (numbers 13, 14, 15, 21, and 22), i.e. the "Robertsonian translocation", may result in an imbalance of Hsa21 in the fetus (Fig. 3). This may be due to a silent translocation in one of the parents or to a *de novo* mutation, with a probability of ~50% for each event (Antonarakis, 1998). In these cases, the extra copy of Hsa21 is not separated ("free") from the others. If one of the parents carries such a mutation, the likelihood of having a child with T21 increases. These subjects are phenotypically normal, but they can produce unbalanced gametes that, when fertilized, result in a monosomic or trisomic zygote.

4. Duplication. A DNA duplication error or a translocation may result in an extra copy of a portion of Hsa21, leading to trisomy of delimited portions of Hsa21 ("partial" or "segmental" T21) (Fig. 4).

These mechanisms lead to different types of T21 (Fig. 5). According to the individuality of the extra copy of Hsa21, we distinguish a **free** form if this copy is physically independent from other DNA molecules or a **translocated** form if Hsa21 genetic material is fused to other chromosomes. When considering cellular distribution of the genetic alteration, we can

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

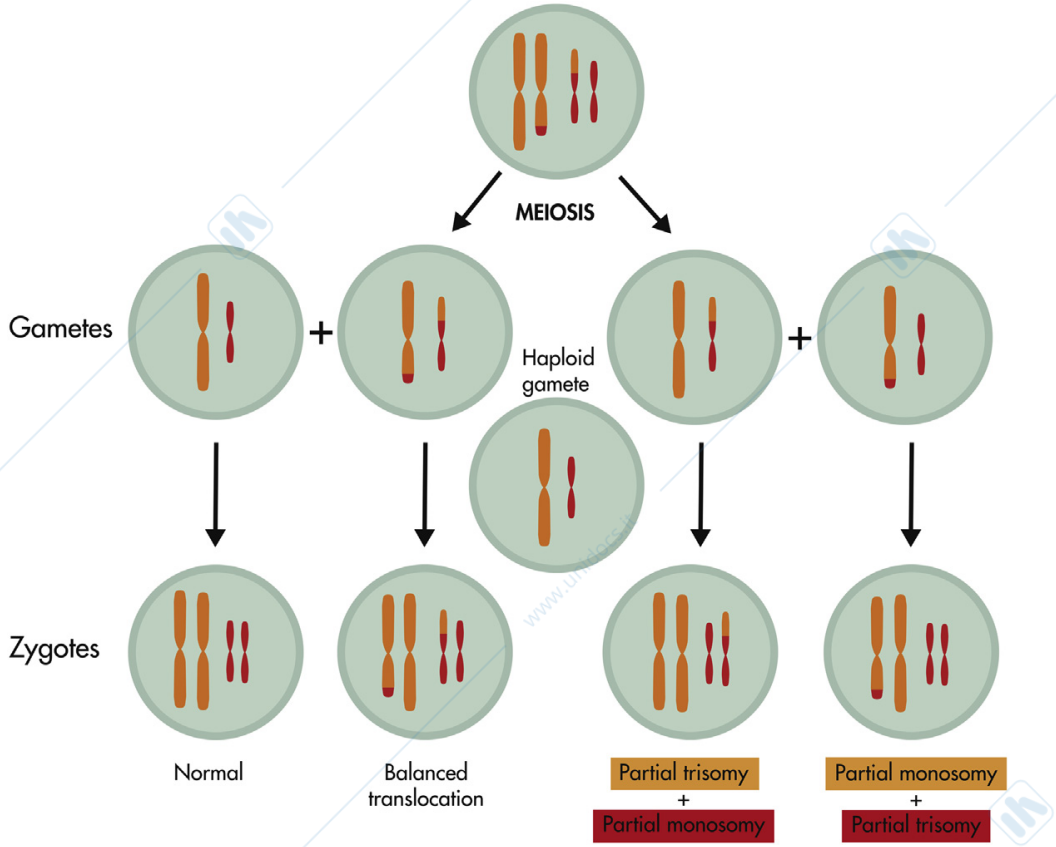


Fig. 4 Partial trisomy caused by unbalanced translocation. *Image courtesy of Marcello Fedi.*

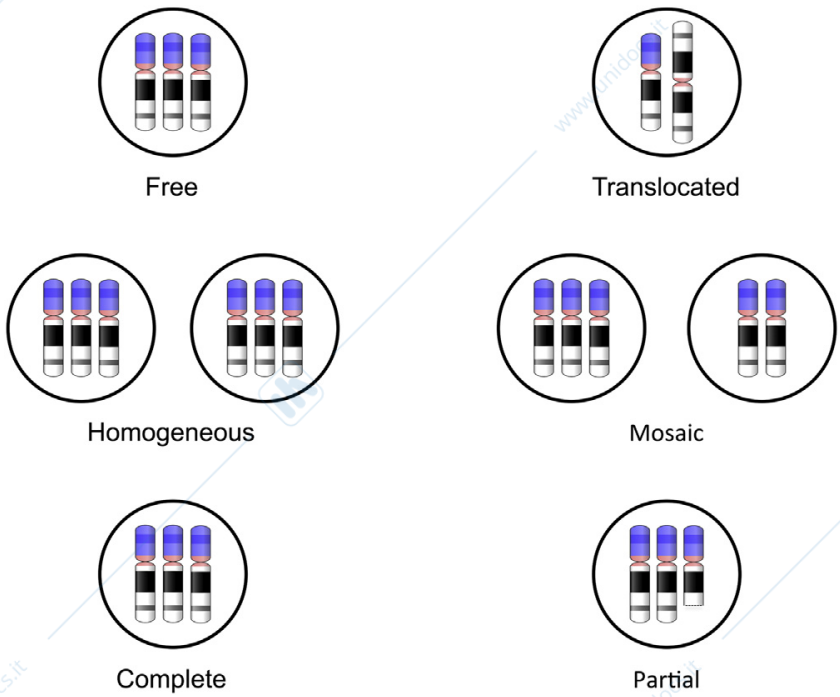


Fig. 5 Different types of T21.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

distinguish a **homogeneous** form if it appears present in all cells of the organism and a **mosaic** form if some cell lineages of the organism carry T21 and some others not, just as mosaic artwork is consisted of different types of pieces brought together. Finally, from the point of view of the integrity of the extra Hsa21 copy, in typical cases the long arm of Hsa21 (21q) is **complete**, i.e. all its genetic material is present, but in rare cases, as shown above, 21q may be incomplete leading to **partial** T21. These classifications, and their combination, will be discussed below.

3.1 Free versus translocated trisomy 21

The most typical genetic basis of DS is free T21, observed in 90%–95% of cases (Devlin & Morrison, 2004a, 2004b; Mutton, Alberman, & Hook, 1996). Due to the biological reasons discussed above and in agreement with an increased probability of errors in advanced maternal age, it has been demonstrated that errors in maternal meiosis I are by far the typical situation. Chromosomal segregation anomalies have been estimated to occur in 68% of cases during meiosis I (~65% maternal, ~3% paternal) and in 28% of cases during meiosis II (~23% maternal, ~5% paternal) (Hassold et al., 1996; Hassold & Hunt, 2001; Karmiloff-Smith et al., 2016). By using DNA polymorphisms as markers to determine the parental origin of the extra Hsa21, it has been determined that the extra Hsa21 is of maternal origin in about 95% of the cases and paternal in only about 5%, contrasting previous reports estimating paternal origin in 20% of cases (Antonarakis, 1991).

Another typical form of DS is also seen when 21q is translocated to another chromosome, in most cases in the form of a fusion of the long arms of two acrocentric chromosomes, i.e. the "Robertsonian translocation" or "centric fusion". DS due to translocation, mostly Robertsonian of the type t(14;21) or t(21;21), is observed in 4% of cases in which mean maternal age was not raised in contrast to free T21 (Mutton et al., 1996). The first description of a such a case was made in 1959 (Turpin & Lejeune, 1961; Turpin, Lejeune, Lafourcade, & Gautier, 1959).

The translocation may occur *de novo* or may be inherited from a phenotypically normal parent carrying the translocated chromosome in the context of a balanced genome of 45 chromosomes (Azevedo Moreira, Damasceno Espirito Santo, & Fernandes Lacerda Carvalho, 2013; Wilch & Morton, 2018). The most likely mechanism for the translocation event is that it occurs before crossing over in meiosis I and is followed by normal segregation in meiosis I and II (Petersen et al., 1991).

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

In a case series, about a third of translocations were inherited, often from the mother (six to seven times more often than the father) (Mutton et al., 1996). The probability to have a child with DS is increased if one parent is a carrier of a Robertsonian translocation. Balanced Robertsonian translocation has a prevalence of 1 in 1000 healthy individuals (Keymolen, Van Berkel, Vorsselmans, Staessen, & Liebaers, 2011), mostly involving chromosomes 13 and 14 or 14 and 21 (Gardner, Grant, & Lisa, 1989).

Hsa21 short arm (21p) is considered genetically empty in practice, containing tandemly repeated ribosomal RNA genes that are redundant in the human genome. Therefore, if the 21q dose is unbalanced, the loss of 21p is consistent with a DS phenotype indistinguishable from the one due to free T21 (Dent, Edwards, & Delhanty, 1963; Gustavson, 1964). On the other hand, if the 21q dose is balanced, there are no clinical consequences (Eggermann & Schwanitz, 2011; Hamerton, 1971; Rethore, 1981).

3.2 Mosaicism

Mosaicism is defined as a condition in which an individual has two or more genetically distinct cell lines that originated from a single zygote (Papavassiliou, Charalsawadi, Rafferty, & Jackson-Cook, 2015) and in DS it was described for the first time in 1961 (Clarke, Edwards, & Smallpeice, 1961). DS in mosaicism is observed in about 1% of cases (Mutton et al., 1996), although according to a recent revision of the subject, mosaic cases are estimated to be more common (up to 5%) than previously recognized (Papavassiliou et al., 2015), and often do not have dysmorphic features.

Besides the mitotic error following fertilization as explained above, a mosaic condition could alternatively originate from an early mitotic error in an embryo with T21 allowing some of the cells to revert to a normal karyotype (Papavassiliou et al., 2015).

While it could be expected that the higher the percentage of trisomic cells in mosaic subjects, the higher grade and/or the broader range of manifestations they display, establishing a genotype-phenotype relationship in mosaic subjects has been hampered by the problem of the topographic distribution of trisomic cells within the organism. In fact, mosaicism may be evaluated only in easily accessible cells, like lymphocytes, buccal cells, or skin biopsy-derived fibroblasts, but poor information may be obtained *in vita* about the percentage of T21 cells in organs like brain, whose function is mainly affected by T21.

As stated by Lejeune in 1969, "Intermingled with all the preceding cases is the cumbersome problem of mosaicism. Every human being is a mosaic

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

due to some mitotic malsegregation in some part of the body; the dividing line between normal and pathologic is a matter of percentage: if 50% of abnormal karyotype is considered as deleterious, what about 40%, 20%, 5%, and what about topographical distributions?" (Lejeune, 1970).

The affirmation that all humans are mosaic to a certain degree, based on the unavoidability of casual errors in mitosis also in genetically normal subjects, was formally shown to be true in 2005 when Rehen and coll. found that 2% of brain neurons in humans have T21 in a *post-mortem* study (Rehen et al., 2005), leading to the concept that genomic mosaicism is a feature of the normal brain (Bushman & Chun, 2013; Hulten et al., 2014). Between the very low physiologic percentage of mosaicism and the almost total T21 cellular condition (98% in typical homogenous T21 because mitotic errors in T21 zygotic karyotype restore a certain percentage of euploid cells (Rehen et al., 2005)), lies the intermediate situation of a variable proportion of T21 cells. Despite the difficulties explained above, a low ratio of T21/normal blood lymphocytes commonly used as an estimation of mosaicism has revealed in practice to be associated with less phenotypic manifestations, higher intelligence quotients (IQs), and better overall survival (Leon, Zou, & Milunsky, 2010). The possibility of low-level mosaicism led to the recommendation to perform further studies, such as SNP microarray analysis, to detect mosaicism in children with features of DS and a normal blood metaphase karyotype (Leon et al., 2010).

Papavissiliou and coll. recently reviewed mosaicism conditions for T21 (Papavassiliou et al., 2015) in 2015. They also observed a significant negative correlation between IQ scores and the proportion of T21 buccal cells (Papavassiliou et al., 2009), with the same trend (although not statistically significant) when lymphocyte was the considered cell type analyzed. The reverse was true for correlation with congenital heart defects (CHDs) and significant for lymphocytes but not for buccal cells (Papavassiliou et al., 2009), although the trend was the same.

3.3 Partial trisomy 21

The presence of three copies of only a delimited segment of Hsa21 was first reported by Ilbery et al. (Ilbery, Lee, & Winn, 1961) as "incomplete trisomy", it may or may not be associated with DS and its natural occurrence is considered to be exceptional or extremely rare (Antonarakis, 1998; Aula, Leisti, & von Koskull, 1973; Pelleri et al., 2016). This condition is commonly referred to as partial, or segmental, T21 (PT21).

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

PT21 may follow a translocation (Fig. 4) or derive from a microduplication of a Hsa21 region (Pelleri et al., 2016).

Regarding PT21, it has been unambiguously shown that some 21q portions are strictly associated with core features of DS while others not, as discussed in detail below in the "Genomics of Down syndrome" Section 7.

Any of the eight possible combinations of the main types of T21 (free or translocated, in a homogeneous or mosaic state, with a complete or partial 21q) could in theory exist, and each of them has actually been described leading to free or translocated Hsa21 with a different distribution in the organism and with a different grade of 21q completeness, some combinations being rare (e.g., Hongell & Airaksinen, 1972; Lejeune, Berger, Vidal, & Rethore, 1965; Maciejewski, Vigneron, Lambert, Bonnet, & Hascoet, 2012; Su et al., 2016; Van Keuren et al., 1989; Zhao et al., 2015) (Table 1). While most standard, and frequent, forms, i.e. free (or translocated) homogeneous complete T21 typically cause DS, the clinical outcome may be highly variable when only a portion of the body cells is affected (mosaicism), depending on the percentage and the function of the T21 cells (Lejeune, 1970) or when only a portion of 21 is trisomic (PT21), depending on the specific gene content of the trisomic region and cellular functions these genes control (Table 1).

4. Signs and symptoms of Down syndrome

Each chromosome acts as an information template for the building of cellular components needed for the correct functions of the cells and of the whole organism. It is expected that possessing three copies of a whole chromosome seriously affects hundreds of functions leading to multiple medical problems in the case of survival and live birth. However, in the case of T21, while the genetic damage might cause, or predispose to, the development of pathologic signs and symptoms in virtually any organ, the most frequent and shared manifestations should be clearly distinguished from the whole phenotypic constellation which is never present in its entirety in a single child thus often times not needing to become a "medicalized" subject.

The spectrum of phenotypic manifestations of DS and of associated comorbidities has been reviewed in detail by several Authors (Epstein, 1989, pp. 291–326; Gardiner et al., 2010; Hickey et al., 2012; Letourneau & Antonarakis, 2012; Megarbane et al., 2009; Roizen & Patterson, 2003; Weijerman & de Winter 2010). The most common features are certainly ID, craniofacial dysmorphisms, and cardiovascular defects (Gardiner et al.,

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Table 1 Combinations of the different types of T21.

Trisomy 21 type	Notes	Percentage	Examples
Free, homogeneous, complete	Standard type	95%	(Lejeune, Gautier, et al., 1959a; Mutton et al., 1996)
Free, in mosaicism, complete	Typical mosaic	1%	(Clarke et al., 1961; Papavassiliou et al., 2015)
Free, homogeneous, partial		Rare	(Ilbery et al., 1961; Su et al., 2016)
Free, in mosaicism, partial		Rare	(Hongell & Airaksinen, 1972; Schnabel et al., 2018)
Translocated, homogeneous, complete	Typical translocated	2%–4%	(Mutton et al., 1996; Turpin et al., 1959)
Translocated, in mosaicism, complete		Rare	(Zhao et al., 2015)
Translocated, homogeneous, partial		Rare	(Lejeune et al., 1965; Maciejewski et al., 2012)
Translocated, in mosaicism, partial		Rare	(Van Keuren et al., 1989)

2010). A seminal 1990 article by Lejeune (Lejeune, 1990) started by describing these signs: "With upward-slanting eyelids, a little nose in a round face, and incompletely chiseled features, DS patients look more like children than the usual child does". Lejeune, who visited thousands of children with T21, recognized that their character is "cheerful and affectionate, they have a special charm easier to cherish than to describe", but also that the condition deprives "the children of the most precious quality afforded by our genetic patrimony, the full power of rational thinking".

ID is present to some degree of severity in virtually all individuals with T21 (Lukowski, Milojevich, & Eales, 2019). Typical area of weakness are psychomotor development, abstract thinking, and speech development whereas affectivity and social skills are conserved (Tolksdorf & Wiedemann, 1981). These features are well recognizable both by "gestalt" impression gained through clinical experience as well as by natural history of developmental trajectory (Fidler, 2005) and neuropsychological testing (Ghezzo

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

et al., 2014; Lanfranchi, Toffanin, Zilli, Panzeri, & Vianello, 2014) and are associates to brain imaging (Menghini, Costanzo, & Vicari, 2011) as well as neural structural correlates (Stagni, Giacomini, Emili, Guidi, & Bartesaghi, 2018). It is common for children with DS to arouse a climate of affective intensity greater than normal so that some Authors have come to speak of a "kindness gene" (Barnicoat et al., 1996) or a "happy personality" (Weisfeld-Adams, Tkachuk, Maclean, Meeks, & Scott, 2016). A study about the self-perception of people with DS, ages 12 and older, found that the overwhelming majority of people with DS were happy with their lives, liked who they were, liked how they looked and expressed love for their families, indicating they live happy and fulfilling lives (Skotko, Levine, & Goldstein, 2011).

Most reliable and discriminative neonatal physical signs of DS include small ears, wide space in between the first and second toe ("sandal gap"), small internipple distance, Brushfield spots, and nuchal skin fold (Weijerman & de Winter 2010) as well as hyperflexibility of joints (Hickey et al., 2012). In addition, also brachycephaly, hypotonia, flat face, upward slant of the eye split, and a single crease across the palms are also reliable and discriminative signs (Weijerman & de Winter 2010), as well as broad, short hands, and an inturned and/or short fifth finger (Jackson, North, & Thomas, 1976). Epicanthic fold is age-dependent, while a low, flat nose bridge and small mouth are very common, although they could be difficult to differentiate (Weijerman & de Winter 2010). Short stature is commonly observed in persons with DS (Hickey et al., 2012; Mircher et al., 2018). While some mothers with DS have been described, fertility is exceptional in men (Pradhan, Dalal, Khan, & Agrawal, 2006).

Another very relevant phenotype is represented by one of several types of CHD, whose prevalence is estimated at 44%–58% worldwide (Weijerman & de Winter 2010).

In addition, many other comorbidities have been observed in the context of a wide phenotypic variability, mainly visual and hearing impairments, obstructive sleep apnea syndrome, increased susceptibility to infections and wheezing airway disorders, and congenital gastrointestinal tract defects such as Hirschprung's disease (Weijerman & de Winter 2010).

Other phenotypes may be observed at a lesser frequency, although still higher in comparison to non-T21 children, such as obesity; autoimmune pathologies such as hypothyroidism, celiac disease (CD), alopecia, and diabetes; atlanto-axial instability; dermatological and dental problems; and hematological disorders such as increased risk of leukemia, in particular

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

megakaryoblastic (AMKL) (Pelleri et al., 2014). While the risk of developing leukemia and testicular cancer is much higher than age-matched controls, women with DS almost never develop breast cancers and other solid tumors in general are rare (Hasle, Clemmensen, & Mikkelsen, 2000; Patja, Pukkala, Sund, Iivanainen, & Kaski, 2006).

Finally, early-onset cognitive decline with neuropathological alterations similar to those observed in the brains of patients with Alzheimer's disease (AD) may occur (Dekker et al., 2018; reviewed in Lott & Head, 2019).

The most discriminating signs and symptoms useful for the clinical diagnosis of DS were determined in the classic works of Hall in 1966 for newborns (Hall, 1966) and of Jackson and coll. in 1976 for children (Jackson et al., 1976). According to Jackson and coll., oblique eyes and a flat nasal bridge have been shown to be the two most discriminating dysmorphic signs of DS; the former was actually chosen by Down to initially name the syndrome itself, and the evidence of the latter by prenatal ultrasound scanning has become widely used to suspect DS.

However, it is assumed that only the demonstration of Hsa21 excess material confirms the diagnosis, with methods that will be briefly reviewed in the next section from the genetic point of view.

5. Genetic diagnosis of trisomy 21

Prenatal diagnosis for chromosomal anomalies is traditionally performed on a sample of fetal cells. Both chorionic villus biopsy (villocentesis, at 10–14 weeks of gestation) and amniocentesis (at 16–20 weeks) are invasive, unpleasant for the mother, expensive (Strachan & Read, 2019), and remain aggravated by a significant rate of adjunctive abortion following the procedures, to the order of 1% (Evans & Andriole, 2008). However, they are usually executed to confirm the suspicion of T21 that may be posed by analysis of the level of biomarkers in maternal serum (alpha-fetoprotein, unconjugated estriol, human chorionic gonadotropin or beta-HCG, inhibin-A and PAPP-A (pregnancy-associated plasma protein A)), while ultrasound scanning may detect non-invasively variations in nuchal translucency (Strachan & Read, 2019). Combined with maternal age, combination of serum and ultrasound markers (including nasal bone development) may estimate a risk (Alldred et al., 2017), without providing a real diagnosis, which remains linked to the demonstration of excess Hsa21 material. More recently, noninvasive prenatal testing (NIPT), exploiting the

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

presence of cell-free DNA molecules in the maternal plasma, approximately 5–10% of which are derived from the placenta, a fetal tissue, has been made available from 9 weeks of gestation onward (Strachan & Read, 2019; Wong & Lo, 2016). An excess of one chromosome may be quantified by DNA sequencing, in comparison with chromosomes present in normal dosage. Although sensitivity and specificity of NIPT is continuously increasing, NIPT is still regarded to date as a screening and not diagnostic test, although reducing the number of cases further requiring invasive methods.

A relevant and often missed point is that even invasive methods, like any other diagnostic method in the real word, cannot be 100% sensitive and specific, in particular in the case of mosaicism and at a greater extent for villocentesis, so that, although there are poor literature in this regard, a fraction of false positive may be expected (e.g., Schuring-Blom, Boer, Knegt, Verjaal, & Leschot, 2002).

Postnatal diagnosis is based on the classical karyotype analysis on lymphocytes following a blood draw. In the case of a discrepancy between clinical appearance and karyotype, analysis at a greater resolution is recommended, such as array comparative genomic hybridization (array-CGH) (Lyle et al., 2009).

The communication of the diagnosis or of its suspicion is a central point to be addressed. This information may have critical consequences for the lives of all subjects involved. Counseling should report all options and parents must not be forced to terminate a pregnancy. In the end, social support may be critical for the perception of the value of the person with disability. As stated by Strachan and Read (2019), "Whatever one's position on this argument, all civilized people must surely agree that society has an obligation to look after people born with disabilities and do whatever is possible to allow them a full life".

6. Human chromosome 21 (Hsa21)

Hsa21 is the smallest human chromosome, although it was initially thought Hsa22 was smaller than Hsa21, thus causing an error in cataloguing human chromosomes in descending order of size in karyograms. It is an acrocentric chromosome of the "G" group. The complete nucleotide sequence of Hsa21 was determined in 2000 by a consortium (Hattori et al., 2000). Following publication of the human genome sequence in 2001, the most updated and refined version of Hsa21 sequence available is

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

released in the reference assembly GRCh38.p12 (https://www.ncbi.nlm.nih.gov/genome/gdv/), annotation 109, including the GenBank record NC_000021.9 displaying 46,709,983 nucleotides. Analysis of NCBI Gene database using GeneBase tool (Piovesan, Caracausi, Antonaros, Pelleri, & Vitale, 2016) shows that, up to January 5, 2019, there are 228 known protein-coding genes catalogued as localized on Hsa21 (retrieved by searching for loci only with a "reviewed" or "validated" gene record including at least one "reviewed" or "validated" messenger RNA or mRNA). Using the same criteria, 106 non-coding RNA (ncRNA) may be found, among which 4 microRNAs (miRNAs): *MIR99AHG* and *MIR548XHG* (21q21.1), *MIR155HG* and *MIR155* (21q21.3), although several tens of other *MIR*-loci are recorded in provisional RefSeq status. Hsa21 is considered among the poorest chromosomes for miRNA-encoding genes per Mb, while being among the richest for long non-coding RNA (lncRNA)-encoding genes (Antonarakis, 2017).

Following parsing by TRAM software 1.3 of human genomic data available up to November 2017, it appears that 1331 expressed sequence tag (EST) clusters, each representing one transcript mapped in a distinct location on Hsa21, are still uncharacterized (Pelleri et al., 2018).

Significantly, the only three autosomal trisomies allowing live births known since 1959–60 are the ones of human chromosomes 13, 18, and 21 (Speicher, Antonarakis, & Motulsky, 2010), and only since 2000 we do know these to be the exact three having the lowest number of genes in absolute in the human genome (320, 269, and 228 protein-coding genes respectively, according to updated NCBI Gene database). Also, according to the ENCODE project results (ENCODE Project Consortium, 2012), Hsa21 appears to be among the chromosomes with the smallest number of functional DNA elements per Mb.

While we need an annotated physical map of Hsa21 that is as complete and accurate as possible, a putatively complete list of Hsa21 functional elements is still lacking. For instance, a large gene extending for more than 100,000 bp on Hsa21 and encoding a protein (CYYR1) conserved in vertebrates (Vitale et al., 2002, 2007) went unrecognized in the gene catalogue originally published along the Hsa21 sequence (Hattori et al., 2000). It should therefore be kept in mind that accurate, individual characterization of individual genomic regions is an ever-going process that could reveal new active loci, provided that a certain previously not studied biological condition is investigated or a new gene discovery method is applied.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

7. The genomics of Down syndrome and the "Down syndrome critical region"

DS represents a condition where there is an excess ("gain") of function, and not a loss of some genetic action. Hsa21 DNA sequence is assumed to be normal in T21, but the third, additional template is expected to lead to overexpression of gene products ("gene overdosage", Fig. 6), in theory at 150% of the normal level. There are several mechanisms by which the 3:2 DNA template dosage for Hsa21 could affect cellular functions, including development (reviewed by Antonarakis, 2017). The excess product could exert a function at a higher rate than the normal or could regulate the expression of other genes. It has also been suggested that haploinsufficient genes (genes whose loss-of-function results in a recognizable phenotype) are also sensitive to an excess gene dosage and are thus good candidates for contributing to some of the T21 phenotypes (Antonarakis, 2017; Conrad & Antonarakis, 2007).

In addition, it should be noted that due to chains of biological relationships, it might not be obligatory for a gene affecting brain functions to be expressed in the brain, provided that it regulates or in any way interacts with a critical gene located anywhere in the genome (Antonarakis, 2017).

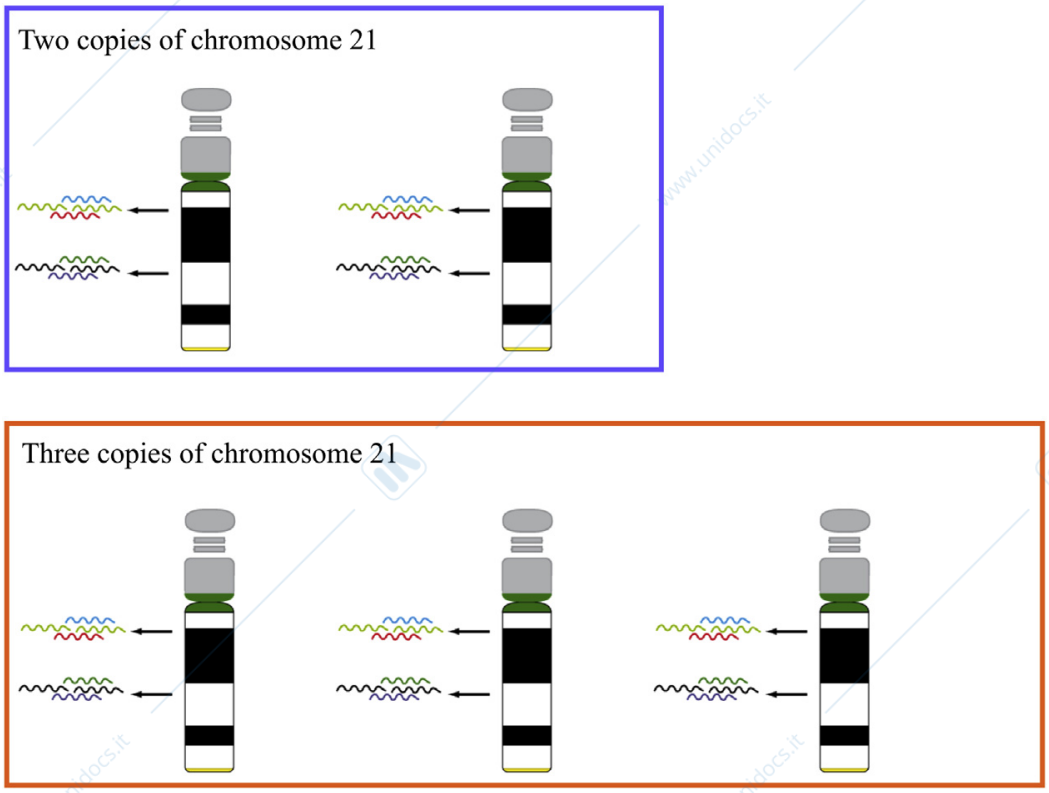


Fig. 6 Gene overdosage associated to trisomy.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

While all the cited mechanisms are expected to lead to some effect at the cellular level, it must again be considered that some mechanisms are expected to be more critical than others. It has indeed been recognized early that studying the relationships between the particular Hsa21 region trisomic in subjects with PT21 and their phenotype offered an invaluable tool to "detect the culprits among so many innocents" (Lejeune, 1990), and on the other hand to show that a 3:2 gene dosage for many loci on Hsa21 may be tolerated. "Surely, most of the genes do not produce harm when in triplicate, because trisomic children would not survive at all. Few of the accelerated reactions are dangerous" (Lejeune, 1990). On the contrary, an extra copy of a small segment of 2.5 Mb due to a microduplication on 21q22, even in mosaicism, has been observed in a child with phenotypic features of DS (Schnabel et al., 2018). In his seminal work published in 1974, Niebuhr reviewed 14 previously described cases with PT21, comparing the extension of retained 21q regions and phenotypes of the affected, concluding that "trisomy of a rather delimited segment on chromosome No. 21 is essential for the development of typical features in Down's syndrome", thus suggesting that "the very distal segment (21q22)" (17.4 Mb) "may be pathogenetic in Down's syndrome" (Niebuhr, 1974).

In the subsequent 20 years, several single cases as well as case series of PT21 were reported. Following the introduction of fluorescence in situ hybridization (FISH) techniques, a more detailed description of the duplicated 21q segments associated to DS was made possible, suggesting in the early 1990s that a delimited sequence interval on Hsa21 overlapped among different cases, thus being a candidate for the basic phenotype characteristic of T21 (Delabar et al., 1993; McCormick et al., 1989; Rahmani et al., 1989). Different studies converged toward a region within 21q22 and restricted up to 0.6 Mb, from 37.7 to 38.3 Mb on Hsa21 (Pelleri et al., 2016). The first time that the term "Down Syndrome critical region" (DSCR) was used appears to have been in 1990 (Rahmani et al., 1990), a term equivalent to "Minimal Chromosomal Region" (MCR) (McCormick et al., 1989), "Down Syndrome minimum critical region" (DCR) (OMIM entry #190685), or "Down Syndrome minimal region" (Shapiro, 1999). The DSCR concept should indicate a region causing the most constant and shared features of DS, rather than a region proposed for being involved in the whole phenotypic spectrum of DS, as pointed out by Rethoré: "The presence in triplicate of subbands 21q22.1 and 21q22.2 suffices to induce the main phenotypic symptoms of the classic syndrome of trisomy 21" (Rethore, 1981). In many cases where the karyotype was reported as normal

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

but the phenotype was strongly suggestive of DS, further studies at a higher resolution revealed PT21 (Pelleri et al., 2016).

Array Comparative Genomic Hybridization (CGH) further allowed fine characterization of cases of DS with segmental trisomy of Hsa21 (Korbel et al., 2009; Lyle et al., 2009). If the child has a certain fragment and has a diagnosis of DS, this fragment is supposed to be necessary for the manifestation of the most common features of the syndrome (ID and some facial phenotypes) and presumably contains the genes that cause them. On the other hand, if the child has a different chromosomal fragment and has not been diagnosed with DS, this fragment may be excluded as responsible for the common features of the syndrome (Fig. 7).

By combining the data obtained from 125 cases reported in the literature from 1973 to 2015/2016, we have built and reported a detailed map of the partial fragments ever described to search for a possibly minimal chromosomal region shared by all children with DS and absent in all children without DS (Pelleri et al., 2016). The novelty of the approach, and consequently of the results, was in the systematic analysis: every report of PT21 was identified by searching old and new archives of biomedical articles; each case description was entirely revised to discuss signs, symptoms and the diagnosis of the child as well as the cytogenetic (chromosome mapping) and molecular data available. In addition, advanced bioinformatic methods were used to combine data obtained by different mapping methods in a coherent overview based on the most updated Hsa21 map. The results

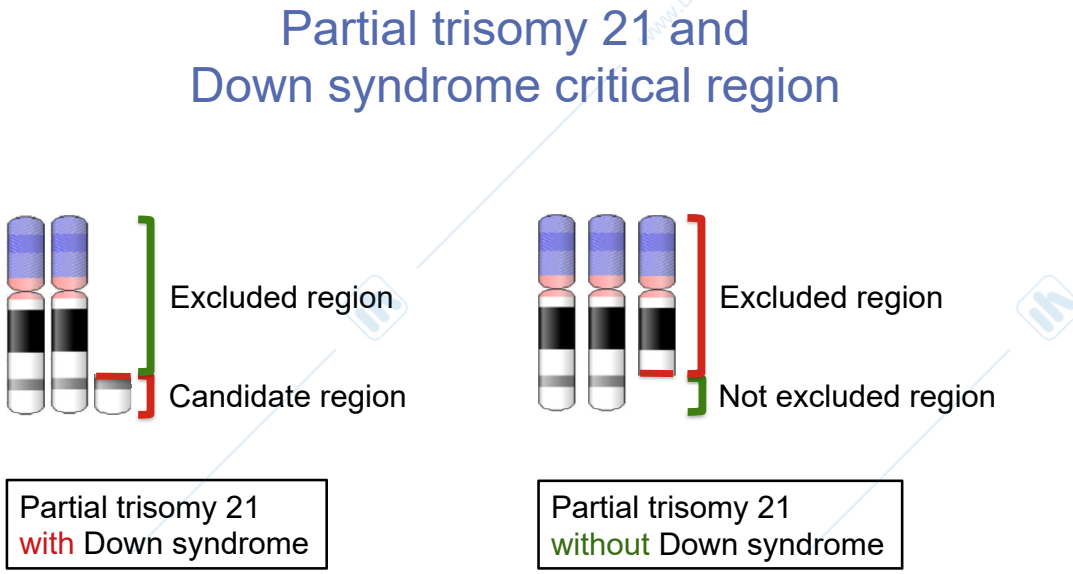


Fig. 7 Partial T21 may or may not be associated with diagnosis of DS, allowing identification or exclusion of Hsa21 regions critical for DS.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

suggest that a region of this type exists on Hsa21, is located within the Hsa21 zone known as 21q22.13 and corresponds only to less than one thousandth of the whole Hsa21, extending from 37,929,229 to 37,963,130 (coordinates from Genome Reference Consortium (GRC) human genome assembly GRCh38, or hg38, December 2013). This region, named "Highly Restricted Down Syndrome Critical Region", or HR-DSCR, contains no known genes in the NCBI reference map and has relevant similarity only with the chimpanzee genome. The proposal of a HR-DSCR paves the way for the identification, in this region, of one or a few currently unknown genetic elements strongly associated to DS, the function of which might become the target for a specific therapy. Recently, through automated annotation of high-throughput RNA sequencing (RNA-seq) data, the Ensembl genome browser (https://www.ensembl.org/, version 95) has mapped an alternative primary RNA transcript of the *KCNJ6* locus, encoding a potassium channel, encompassing the HR-DSCR.

8. Genotype-phenotype correlation in Down syndrome

When in 1988 the beginning of a more systematic characterization of the human genome by DNA sequencing appeared not to be impossible, Lejeune highlighted the need that the forthcoming effort should remain tightly linked to the patient, stating, "The tedious and laborious comparison of the clinical data and of the DNA deciphering is, currently, the starting point of any pathogenic scheme" (Lejeune, 1988, pp. 1–18).

Following actual sequencing of Hsa21 and mapping of Hsa21 genes, several genes have been candidated for DS-related phenotypes on the basis of product function, in particular 16 loci with a role in energy and reactive oxygen species metabolism, including *SOD1* (Superoxide dismutase 1, soluble); 9 loci affecting brain development, neuronal loss, and Alzheimer's type neuropathology, including *SIM2* (Single-minded homolog 2 (Drosophila)), *DIRK1A* (Dual-specificity tyrosine-(Y)-phosphorylation kinase 1A) and *APP* (Amyloid beta (A4) precursor); and 6 loci with a role in folate and methyl group metabolism, including *CBS* (cystathionine-beta-synthase) (Roizen & Patterson, 2003).

While it remains difficult to establish a clear model of molecular pathogenesis for DS and for distinct features associated with DS, a more consistent genotype-phenotype relationship has emerged from structural data derived

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Table 2 Genotype-phenotype correlation examples.

Phenotype	Hsa21 region	Extension (hg38 coordinates)	Known genes	References
Diagnosis of DS	21q22.13	0.034 Mb (37,929,229 −37,963,130)	*KCNJ6* (RNA isoform in Ensembl genome browser version 95)	(Pelleri et al., 2016)
CHD	21q22.2	0.96 Mb (40,236,381 −41,200,000)	*DSCAM*, *BACE2*, *PLAC4* ncRNA: *DSCAM-AS1*, *DSCAM-IT1*, *LINC00323*, *MIR3197*	(Pelleri et al., 2017)
AMKL/TMD	21q22.12-q22.3	8.3 Mb	*RUNX1*, *ERG*, *ETS2*	(Korbel et al., 2009)
Hirschsprung disease	21q22.11-q22.3	12.76 Mb	*ATP50*, *CLIC6*, *ETS2*, *HMGN1*, *SH3BGR*, *DSCAM*	(Korbel et al., 2009)

TMD, transient myeloproliferative disease. Other abbreviations are explained in the text.

from the study of PT21. Besides the association of a critical region to the diagnosis itself of DS (explained in the previous section), a PT21 map built from real cases has revealed to be most useful to map distinct phenotypes on different Hsa21 regions (Korbel et al., 2009; Lyle et al., 2009). Following an updated meta-analysis we have proposed mapping, and relative candidate genes, for DS-associated CHD (Pelleri et al., 2017). We summarize current genotype-phenotype knowledge for some phenotypes of DS in Table 2.

Another approach to identify specific genomic regions involved in DS has been searching for specific genomic domains showing a dysregulation pattern and named Gene Expression Dysregulated Domains (GEDDs) (Antonarakis, 2017; Letourneau et al., 2014). GEDDs were initially found by elaboration of RNA-Seq data deriving from monozygotic twins discordant for DS, comparing regional gene expression in DS and euploid cells. An independent reanalysis of the same data failed to find GEDDs, questioning the validity of this concept (Do, Mobley, & Singhal, 2015).

Finally, while the role of specific chromosomal regions and Hsa21 loci could be shared by different subjects with DS, sequence variants are an adjunctive factor to be considered, possibly justifying interindividual differences at both nucleotide and submicroscopic chromosomal levels (Antonarakis, 2017).

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Several reviews on the genotype-phenotype correlation in DS are available (Antonarakis, 2017; Patterson, 2009; Patterson & Costa, 2005).

9. Beyond genomics: epigenomics, transcriptomics, proteomics and metabolomics of Down syndrome

Apart from studies on individual human genes typical before 2000, it later became feasible to study in systematic form the expression of the information contained in the whole genome in terms of its modulation by epigenetic mechanisms, of its expression as a transcriptome, the pool of RNA transcripts in a cell type, up to its action in the form of a proteome and the effects of the protein action on whole sets of metabolites (metabolome).

The relevance of epigenetic mechanisms in DS has been recognized in the last years (Sanchez-Mut, Huertas, & Esteller, 2012). The "epigenetic clock" has recently been used as a quantitative molecular marker of aging, finding that T21 significantly increases the age of blood and brain tissue (Horvath et al., 2015). In particular, **epigenome** studies in DS have identified a specific DNA methylation signature in blood cells (Bacalini et al., 2015) and have shown that T21 alters DNA methylation in both parent-of-origin-dependent and -independent manners (Alves da Silva et al., 2016).

The study of the global gene expression profile (**transcriptome**) has been used by several Authors to verify if the 3:2 ratio of DNA templates for Hsa21 genetic information actually results in a 3:2 excess of the corresponding RNA transcripts. Indeed, following initial reports using expression microarray technology (Giannone et al., 2004), it was shown, until recently, that the expected overexpression of Hsa21 loci is found in all cells and tissues studied in this regard (Letourneau & Antonarakis, 2012). We have conducted a systematic meta-analysis integrating all the available gene expression profiles obtained by expression microarrays and useful to compare DS and normal tissues, generating quantitative differential transcriptomic maps for tens of thousands of transcripts from normal brain, lymphoblastoid cell lines, blood cells, fibroblasts, thymus, and induced pluripotent stem cells, respectively (Pelleri et al., 2018). The recent transcriptome analysis of genetically matched human iPSC of the same subject, disomic or trisomic for Hsa21, maintained as iPSCs or differentiated in cortical neurons, confirmed specific alteration of transcriptome in DS (Gonzales et al., 2018). The availability of the results of the ENCODE project, providing a systematic analysis of functional elements in the human genome, will facilitate the interpretation of Hsa21 sequence in functional terms (ENCODE Project Consortium, 2012).

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

A few studies of **proteome** have been also conducted in DS, confirming the primary excess information template results in recognizable consequences up to the protein expression profile. It has been possible to identify a plasma N-glycome signature typical of DS (Borelli et al., 2015), as well as protein damage by the lipid peroxidation by-product 4-hydroxy-2-nonenal (HNE) (reviewed by Barone, Head, Butterfield, & Perluigi, 2017).

Regarding **metabolome**, we have recently reported for the first time an analysis of the Nuclear Magnetic Resonance (NMR)-detectable part of the metabolome in plasma and urine samples from individuals with DS and typically developing individuals selected among their siblings, showing that metabolomic profiles allow a clear discrimination (up to of 80% accuracy) between the two groups. A significant alteration for some individual metabolites, several produced at the beginning or during the Krebs cycle, was also observed and was consistent with the 3:2 gene dosage model (Caracausi et al., 2018), while accounting for sex, age and fasting state did not significantly affect the results. Metabolome approaches could reveal critical targets of intervention, as it has happened for several metabolic genetic diseases. To once again cite Lejeune, "Phenylketonuria, galactosemia, vitamin B6 dependent homocystinuria, to take few examples, can be properly handled and the children protected against mental deficiency. Who could believe that during the coming years no new progress will be achieved?" (Lejeune, 1980). Lejeune stressed the relevance of the one-carbon metabolism in DS (Gueant et al., 2003; Lejeune et al., 1986), and we have recently reported differences in distinct folic acid derivates in rescuing methotrexate toxicity on *in vitro* cultured fibroblasts (Vitale et al., 2019).

Finally, gut **microbiome,** the whole set of microbial organisms naturally residing in the intestine, has also been found to be specifically altered in DS (Biagi et al., 2014).

10. Human cellular models of trisomy 21

The usefulness of *in vitro* study of T21 cells was recognized early on and, over time, different systems have been devised to understand cellular and molecular biology of T21 cells. Most commonly used cells have been primary cells obtained by subjects with DS, T21 cell lines established *in vitro*.

First, cultured child and adult skin fibroblasts have been extensively used over decades to model cellular and molecular features of T21, such as cell proliferation, enzyme activity, gene expression, sensitivity to radiation, and DNA repair, as reviewed (Vitale et al., 2019), as well as the rescue effect

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

on methotrexate toxicity mediated by folic acid and some of its derivatives (Vitale et al., 2019). Moreover, lymphocytes (e.g., Agarwal et al., 1970; Lejeune et al., 1986) and lymphoblastoid cell lines (e.g., Ait Yahya-Graison et al., 2007; Granese et al., 2013; Prandini et al., 2007; Spellman, Ahmed, Dubach, & Gardiner, 2013; Taylor et al., 1988) have been classically used for the same goal. These models are useful to study general characteristics of T21 cells, or specific of the relative cell types, but lack the proximity to neural cell types likely to be involved in the mechanisms leading to ID.

More recently, an induced pluripotent stem cell (iPSC) model has emerged, that may be induced to differentiate toward various cell types and tissues (reviewed in Antonarakis, 2017), thus elucidating the role of Hsa21 genes by comparing effects of their trisomic or disomic (control) state in a certain cell type. Human T21 iPSC lines have been obtained by fibroblasts of subjects with DS or by amniocytes, along with relative control euploid cells. The first report about T21 iPSCs was published in 2008 (Park et al., 2008) and has been followed by reports studying T21 iPSCs differentiated in neurons (e.g., Shi et al., 2012; reviewed in Zhao & Bhattacharyya, 2018), hematopoietic progenitors (e.g., Maclean et al., 2012), and recently astrocytes (Araujo et al., 2018). iPSCs have been shown to be useful for the study, among others, of genetic and neural developmental features of DS etiology (Briggs et al., 2013). Possible pitfalls of the iPSC model include the possibility that they suffer deep genotype rearrangements (Sobol et al., 2015) or spontaneous chromosome loss (Li et al., 2012). Due to variability of iPSC lines produced even from a single sample, it has been suggested that more than one such iPSC line per sample should be produced (reviewed in Antonarakis, 2017).

Finally, the obtainment *in vitro* of T21 cells with a copy of *XIST* transferred through zinc finger nucleases (ZNFs) in one Hsa21 resulting in an inactivation of one extra copy of the chromosome (Jiang et al., 2013) paves the way for the study of the differences between T21 and functionally euploid cells of the same line. This possibility was first foreseen in a therapeutic perspective by Lejeune when in 1977 he proposed "turning off the extra chromosome by some kind of induced inactivation (like the lyonization of supernumerary X chromosomes)" (Lejeune, 1977), when the basic mechanism of this inactivation was still unknown. The discovery of *XIST* as the gene able to switch off one of the two X chromosomes in female cells through production of a large noncoding RNA responsible for X heterochromatinization (lyonization) (Clemson, McNeil, Willard, & Lawrence, 1996) made this option available in an iPSC model by the same group (Jiang

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

et al., 2013). While there are also some limits in this model (5% of the Hsa21 genes remain active and they could maintain a possibly significant part of the phenotype or the two homologous chromosomes remaining active might be the ones from the same parent thus leading to uniparental disomy), it has recently been used to demonstrate normalization of hemopoietic defects *in vitro* following Hsa21 inactivation through *XIST* (Chiang, Jiang, Newburger, & Lawrence, 2018).

11. Animal models of human trisomy 21

The most useful animal model would in theory be a trisomic chimpanzee, being that *Pan troglodytes* is the animal with the genome, anatomy, and physiology most similar to humans. In addition, chimpanzees appear to date to be the only animal with a naturally occurring aneuploidy equivalent to human T21, with a case being described in 1969 (McClure, Belden, Pieper, & Jacobson, 1969) and a new case recently in 2017 (Hirata, Hirai, Nogami, Morimura, & Udono, 2017). However, due to difficulty and costs of chimpanzee maintenance, they are of very limited use in DS research as well as in research in general.

The basis for the proposal of mice as a model organism for DS was the fact that both mice and humans belong to the mammal class and there is the conservation of many protein coding genes between Hsa21 and murine (Mmu) chromosomes 10, 16, and 17 (Edgin, Mason, Spano, Fernandez, & Nadel, 2012). The first, and still most used, mouse model for T21 was obtained by Davisson in 1990 establishing Ts65Dn, a murine strain with a segmental trisomy of Mmu16, limited to the part of the chromosome homolog to Hsa21 (Davisson, Schmidt, & Akeson, 1990). Since then, many other murine models have been proposed and they are reviewed in several articles (Edgin et al., 2012; Gotti, Caricati, & Panzica, 2011; Kleschevnikov, Belichenko, Salehi, & Wu, 2012; Liu et al., 2011; Rueda, Florez, & Martinez-Cue, 2012; Yamakawa, 2012). One of the recent strains has rearranged Mmu16, Mmu17, and Mmu10 in order to possibly include trisomy of the full spectrum of Hsa21 genes with orthologs on mice (Yu et al., 2010).

Nevertheless, mice mimicking DS remain difficult to grow and breed, and great differences still exist between the two species, hampering a direct translation of evidence derived from the murine DS model to humans (Zhao & Bhattacharyya, 2018). For instance, memantine was reported to be effective in DS model mice but it was shown to be ineffective in adults with DS

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

(Hanney et al., 2012) and to ameliorate some abilities in young adults with DS (Boada et al., 2012). Therefore, while mouse models for DS appear to be a very common approach to the study of T21, attention should be brought to the known multiple limits of these models (Nelson & Gibbs, 2004), in particular when comparing the impairment of superior functions exclusive of human intelligence (Lejeune, 1966) such as abstraction and language, specifically damaged in DS, to alterations of mouse behavior (Nuffield Council on Bioethics, 2002). Finally, although murine *Kcnj6* and *Kcnj15* are syntenic with the human orthologs, no relevant homology may be found in the mouse genome with the HR-DSCR sequence itself that is located between *KCNJ6* and *KCNJ15* in humans (Pelleri et al., 2016).

12. Integrated study and perspectives of the genetics and genomics of Down syndrome

The first observation of T21 in children with DS was made in 1959, only six years after the elucidation of the molecular structure of DNA. The six decades up to 2019 have since then seen the remarkable development of molecular, human and medical genetics, from the elucidation of the basic structure of the gene up to the generation of systematic functional genomics data. While in the '90s Lejeune could count on a few genes mapped on Hsa21 to elaborate his pathogenic models of DS, there is now the unprecedented possibility to integrate sequence, functional, and clinical data in order to identify rational targets for an intervention.

Lejeune clearly remarked the need for the integration among different biomedical fields, stating, "It seems that only when clinical symptoms, gene mapping and biochemical disturbances are correlated, the first possibility of understanding the mechanism of mental deficiency will really be open" (Lejeune, 1977).

Study of DS has focused on different aspects of the condition (Reeves et al., 2019). However, in recent years the notion that DS is essentially a genetic disease with a well identified marker, Hsa21, has not been fully exploited to concentrate the research for a treatment, in particular of ID, on functional elements of Hsa21 strictly associated to the core features of DS. We hypothesize that the accumulation of data from any experimental and clinical source will hopefully lead, through the understanding of critical links between genotype and phenotypes in DS, to a rational background for an effective treatment, in particular of ID (Strippoli et al., 2013). The words of Lejeune early after the discovery of T21 still remain true: "The

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

biochemical disorders resulting from this trisomy still remain to be discovered, and their highlighting alone could serve as a basis for a pathogenic explanation and, perhaps, a therapeutic attempt" (Lejeune, Turpin, et al., 1959b).

Therefore, in 2019, 60 years after the discovery of T21 and 50 years after the first moon landing (still a source of inspiration today when trying something that seems impossible (Chaikin, 2019)), let us conclude with the Lejeune's positive hypothesis that we will reach this goal one day. "We will beat this disease. It's inconceivable that we won't. It will take much less intellectual effort than sending a man to the Moon" (Lejeune-Gaymard, 2012).

Acknowledgments

This work is dedicated to Professor Jérôme Lejeune (1926–94), who taught us to hate the disease and to love the patient and whose scientific thought is guiding our study, and to his fellow and teacher to some of us Professor Maria Zannotti, who brought back research on DS in Bologna in the late '60's and is still advising us on the subject. We wish to sincerely thank all children, families, colleagues, and students participating in our study on trisomy 21. Our heartfelt thanks to all foundations, associations, families, companies, and friends supporting our research by donations and by promoting fund raising initiatives. Special thanks to Marcello Fedi for providing Figures 1, 2, 3, and 4. We are very grateful to Kirsten Welter for her kind, expert, and timely revision of the manuscript.

References

Agarwal, S. S., Blumberg, B. S., Gerstley, B. J., London, W. T., Sutnick, A. I., & Loeb, L. A. (1970). DNA polymerase activity as an index of lymphocyte stimulation: Studies in Down's syndrome. *Journal of Clinical Investigation, 49*(1), 161–169.

Ait Yahya-Graison, E., Aubert, J., Dauphinot, L., Rivals, I., Prieur, M., Golfier, G., … Potier, M. C. (2007). Classification of human chromosome 21 gene-expression variations in Down syndrome: Impact on disease phenotypes. *The American Journal of Human Genetics, 81*(3), 475–491.

Alldred, S. K., Takwoingi, Y., Guo, B., Pennant, M., Deeks, J. J., Neilson, J. P., & Alfirevic, Z. (2017). First trimester ultrasound tests alone or in combination with first trimester serum tests for Down's syndrome screening. *Cochrane Database of Systematic Reviews, 3*, Cd012600.

Allen, G., Benda, C. E., Book, J. A., Carter, C. O., Ford, C. E., Chu, E. H., … Yannet, H. (1961). Mongolism. *American Journal of Human Genetics, 13*(4), 426.

Alves da Silva, A. F., Machado, F. B., Pavarino, E. C., Biselli-Perico, J. M., Zampieri, B. L., da Silva Francisco Junior, R., … Medina-Acosta, E. (2016). Trisomy 21 alters DNA methylation in parent-of-origin-dependent and -independent manners. *PLos One, 11*(4), e0154108.

Antonarakis, S. E. (1991). Parental origin of the extra chromosome in trisomy 21 as indicated by analysis of DNA polymorphisms. Down Syndrome Collaborative Group. *The New England Journal of Medicine, 324*(13), 872–876.

Antonarakis, S. E. (1998). 10 years of Genomics, chromosome 21, and Down syndrome. *Genomics, 51*(1), 1–16.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Antonarakis, S. E. (2017). Down syndrome and the complexity of genome dosage imbalance. *Nature Reviews. Genetics, 18*(3), 147–163.

Araujo, B. H. S., Kaid, C., De Souza, J. S., Gomes da Silva, S., Goulart, E., Caires, L. C. J, … Cavalheiro, E. A. (2018). Down syndrome iPSC-derived astrocytes impair neuronal synaptogenesis and the mTOR pathway in vitro. *Molecular Neurobiology, 55*(7), 5962–5975.

Aula, P., Leisti, J., & von Koskull, H. (1973). Partial trisomy 21. *Clinical Genetics, 4*(3), 241–251.

Azevedo Moreira, L. M., Damasceno Espirito Santo, L., & Fernandes Lacerda Carvalho, A. (2013). Down syndrome due to rare inherited 15/21 robertsonian translocation: Genetics and reproductive counseling. *International Medical Review on Down Syndrome, 17*(3), 36–38.

Bacalini, M. G., Gentilini, D., Boattini, A., Giampieri, E., Pirazzini, C., Giuliani, C, … Garagnani, P. (2015). Identification of a DNA methylation signature in blood cells from persons with Down Syndrome. *Aging, 7*(2), 82–96.

Barnicoat, A. J., Bonneau, J. L., Boyd, E., Docherty, Z., Fennell, S. J., Huret, J. L., … Tolmie, J. (1996). Down syndrome with partial duplication and del (21) syndrome: Study protocol and call for collaboration. Study I: Clinical assessment. *Clinical Genetics, 49*(1), 20–27.

Barone, E., Head, E., Butterfield, D. A., & Perluigi, M. (2017). HNE-modified proteins in Down syndrome: Involvement in development of Alzheimer disease neuropathology. *Free Radical Biology and Medicine, 111*, 262–269.

Biagi, E., Candela, M., Centanni, M., Consolandi, C., Rampelli, S., Turroni, S., … Brigidi, P. (2014). Gut microbiome in Down syndrome. *PLos One, 9*(11), e112023.

Boada, R., Hutaff-Lee, C., Schrader, A., Weitzenkamp, D., Benke, T. A., Goldson, E. J, & Costa, A. C. (2012). Antagonism of NMDA receptors as a potential treatment for Down syndrome: A pilot randomized controlled trial. *Translationalational Psychiatry [electronic Resource], 2*, e141.

Borelli, V., Vanhooren, V., Lonardi, E., Reiding, K. R., Capri, M., Libert, C., … Wuhrer, M. (2015). Plasma N-glycome signature of Down syndrome. *Journal of Proteome Research, 14*(10), 4232–4245.

Briggs, J. A., Sun, J., Shepherd, J., Ovchinnikov, D. A., Chung, T. L., Nayler, S. P., … Wolvetang, E. J. (2013). Integration-free induced pluripotent stem cells model genetic and neural developmental features of Down syndrome etiology. *Stem Cells, 31*(3), 467–478.

Bushman, D. M., & Chun, J. (2013). The genomically mosaic brain: Aneuploidy and more in neural diversity and disease. *Seminars in Cell & Developmental Biology, 24*(4), 357–369.

Caracausi, M., Ghini, V., Locatelli, C., Mericio, M., Piovesan, A., Antonaros, F., … Cocchi, G. (2018). Plasma and urinary metabolomic profiles of Down syndrome correlate with alteration of mitochondrial metabolism. *Scientific Reports, 8*(1), 2977.

Chaikin, A. (2019). *A man on the moon*. London, England: Penguin Books.

Chang, Y. L., Yi, W. P., Chao, A. S., Chen, K. J., Cheng, P. J., Wang, T. H, & Chang, S. D. (2017). Monozygotic twins discordant for trisomy 21: Discussion of etiological events involved. *Taiwanese Journal of Obstetrics & Gynecology, 56*(5), 681–685.

Chiang, J. C., Jiang, J., Newburger, P. E., & Lawrence, J. B. (2018). Trisomy silencing by XIST normalizes Down syndrome cell pathogenesis demonstrated for hematopoietic defects in vitro. *Nature Communications, 9*(1), 5180.

Choi, S. A., Ko, J. M., Shin, C. H., Yang, S. W., Choi, J. S., & Oh, S. K. (2013). Monozygotic twin discordant for Down syndrome: Mos 47,XX,+21/46,XX and 46,XX. *European Journal of Pediatrics, 172*(8), 1117–1120.

Citterio, C. (1964). Mongoloid idiocy in 1 of 2 monozygote twins. *Rivista di Neurobiologia, 10*, 383–389.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Clarke, C. M., Edwards, J. H., & Smallpeice, V. (1961). 21-trisomy/normal mosaicism in an intelligent child with some mongoloid characters. *Lancet, 1*(7185), 1028–1030.

Clemson, C. M., McNeil, J. A., Willard, H. F., & Lawrence, J. B. (1996). XIST RNA paints the inactive X chromosome at interphase: Evidence for a novel RNA involved in nuclear/chromosome structure. *Journal of Cell Biology, 132*(3), 259–275.

Cocchi, G., Gualdi, S., Bower, C., Halliday, J., Jonsson, B., Myrelid, A., … Anneren, G. (2010). International trends of Down syndrome 1993–2004: Births in relation to maternal age and terminations of pregnancies. *Birth Defects Research Part A: Clinical and Molecular Teratology, 88*(6), 474–479.

Conrad, B., & Antonarakis, S. E. (2007). Gene duplication: A drive for phenotypic diversity and cause of human disease. *Annual Review of Genomics and Human Genetics, 8*, 17–35.

Coppedè, F. (2015). The genetics of folate metabolism and maternal risk of birth of a child with Down syndrome and associated congenital heart defects. *Frontiers in Genetics, 6*, 223.

Coppedè, F. (2016). Risk factors for Down syndrome. *Archives of Toxicology, 90*(12), 2917–2929.

Dahoun, S., Gagos, S., Gagnebin, M., Gehrig, C., Burgi, C., Simon, F., … Blouin, J. L. (2008). Monozygotic twins discordant for trisomy 21 and maternal 21q inheritance: A complex series of events. *American Journal of Medical Genetics, Part A, 146a*(16), 2086–2093.

Davisson, M. T., Schmidt, C., & Akeson, E. C. (1990). Segmental trisomy of murine chromosome 16: A new model system for studying Down syndrome. *Progress in Clinical & Biological Research, 360*, 263–280.

de Graaf, G., Buckley, F., & Skotko, B. G. (2016). Live births, natural losses, and elective terminations with Down syndrome in Massachusetts. *Genetics in Medicine, 18*(5), 459–466.

Dekaban, A. (1965). Twins, probably monozygotic: One mongoloid with 48 chromosomes, the other normal. *Cytogenetics, 4*(4), 227–239.

Dekker, A. D., Sacco, S., Carfi, A., Benejam, B., Vermeiren, Y., Beugelsdijk, G., … De Deyn, P. P. (2018). The behavioral and psychological symptoms of dementia in down syndrome (BPSD-DS) scale: Comprehensive assessment of psychopathology in down syndrome. *Journal of Alzheimer's Disease, 63*(2), 797–819.

Delabar, J. M., Theophile, D., Rahmani, Z., Chettouh, Z., Blouin, J. L., Prieur, M., … Sinet, P. M. (1993). Molecular mapping of twenty-four features of Down syndrome on chromosome 21. *European Journal of Human Genetics: EJHG, 1*(2), 114–124.

Dent, T., Edwards, J. H., & Delhanty, J. D. (1963). A partial mongol. *Lancet, 2*(7306), 484–487.

Devlin, L., & Morrison, P. J. (2004a). Accuracy of the clinical diagnosis of Down syndrome. *The Ulster Medical Journal, 73*(1), 4–12.

Devlin, L., & Morrison, P. J. (2004b). Mosaic Down's syndrome prevalence in a complete population study. *Archives of Disease in Childhood, 89*(12), 1177–1178.

Do, L. H., Mobley, W. C., & Singhal, N. (2015). Questioned validity of gene expression dysregulated domains in down's syndrome. *F1000Research, 4*, 269.

Down, J. L. (1866). Observations on an ethnic classification of idiots. *London Hospital Reports, 3*, 259–262.

Edgin, J. O., Mason, G. M., Spano, G., Fernandez, A., & Nadel, L. (2012). Human and mouse model cognitive phenotypes in down syndrome: Implications for assessment. *Progress in Brain Research, 197*, 123–151.

Eggermann, T., & Schwanitz, G. (2011). Genetics of down syndrome. In S. Dey (Ed.), *Genetics and etiology of Down syndrome* (pp. 3–22). Rijeka, Croatia: InTech.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57–74.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Engler, M. (1945). The causation of mongolism and its prognosis. *American Journal of Mental Deficiency, 50*, 27–34.

Epstein, C. J. (1989). Down syndrome, trisomy 21. In *Metabolic basis of inherited disease*. New York, NY: McGraw-Hill.

Esquirol, E. (1838). *Des maladies mentales considérées sous les rapports médical, hygiénique et médico-légal: J.-B. Baillière*.

Evans, M. I., & Andriole, S. (2008). Chorionic villus sampling and amniocentesis in 2008. *Current Opinion in Obstetrics & Gynecology, 20*(2), 164–168.

Fidler, D. J. (2005). The emerging down syndrome behavioral phenotype in early childhood - implications for practice. *Infants & Young Children, 18*, 86–103.

Fraser, J., & Mitchell, A. (1876). Kalmuck idiocy: Report of a case with autopsy with notes on 62 cases. *Journal of Mental Science, 22*(161–162), 169–179.

Gardiner, K., Herault, Y., Lott, I. T., Antonarakis, S. E., Reeves, R. H., & Dierssen, M. (2010). Down syndrome: From understanding the neurobiology to therapy. *Journal of Neuroscience, 30*(45), 14943–14945.

Gardner, R. J. M., Grant, R. S., & Lisa, G. S. (1989). *Chromosome abnormalities and genetic counseling*. Oxford, UK.

Ghezzo, A., Salvioli, S., Solimando, M. C., Palmieri, A., Chiostergi, C., Scurti, M., … Franceschi, C. (2014). Age-related changes of adaptive and neuropsychological features in persons with Down syndrome. *PLos One, 9*(11), e113111.

Giannone, S., Strippoli, P., Vitale, L., Casadei, R., Canaider, S., Lenzi, L., … Zannotti, M. (2004). Gene expression profile analysis in human T lymphocytes from patients with Down syndrome. *Annals of Human Genetics, 68*(Pt 6), 546–554.

Gonzales, P. K., Roberts, C. M., Fonte, V., Jacobsen, C., Stein, G. H., & Link, C. D. (2018). Transcriptome analysis of genetically matched human induced pluripotent stem cells disomic or trisomic for chromosome 21. *PLos One, 13*(3), e0194581.

Gotti, S., Caricati, E., & Panzica, G. (2011). Alterations of brain circuits in Down syndrome murine models. *Journal of Chemical Neuroanatomy, 42*(4), 317–326.

Granese, B., Scala, I., Spatuzza, C., Valentino, A., Coletta, M., Vacca, R. A., … Andria, G. (2013). Validation of microarray data in human lymphoblasts shows a role of the ubiquitin-proteasome system and NF-kB in the pathogenesis of Down syndrome. *BMC Medical Genomics, 6*, 24.

Gueant, J. L., Gueant-Rodriguez, R. M., Anello, G., Bosco, P., Brunaud, L., Romano, C., … Namour, B. (2003). Genetic determinants of folate and vitamin B12 metabolism: A common pathway in neural tube defect and Down syndrome? *Clinical Chemistry and Laboratory Medicine, 41*(11), 1473–1477.

Gustavson, K. H. (1964). *Down's syndrome: A clinical and cytogenetic investigation*. Uppsala: Almqvist & Wiksell.

Hall, B. (1966). Mongolism in newborn infants. An examination of the criteria for recognition and some speculations on the pathogenic activity of the chromosomal abnormality. *Clinical Pediatrics, 5*(1), 4–12.

Hamerton, J. L. (1971). Banding patterns of metaphase chromosomes in Down's syndrome. *Lancet, 2*(7726), 709.

Hanney, M., Prasher, V., Williams, N., Jones, E. L., Aarsland, D., Corbett, A., … Ballard, C. (2012). Memantine for dementia in adults older than 40 years with down's syndrome (MEADOWS): A randomised, double-blind, placebo-controlled trial. *Lancet, 379*(9815), 528–536.

Hartshorne, G. M., Lyrakou, S., Hamoda, H., Oloto, E., & Ghafari, F. (2009). Oogenesis and cell death in human prenatal ovaries: What are the criteria for oocyte selection? *Molecular Human Reproduction, 15*(12), 805–819.

Hasle, H., Clemmensen, I. H., & Mikkelsen, M. (2000). Risks of leukaemia and solid tumours in individuals with Down's syndrome. *Lancet, 355*(9199), 165–169.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Hassold, T., & Hunt, P. (2001). To err (meiotically) is human: The genesis of human aneuploidy. *Nature Reviews. Genetics, 2*(4), 280–291.

Hassold, T., Abruzzo, M., Adkins, K., Griffin, D., Merrill, M., Millie, E., … Zaragoza, M. (1996). Human aneuploidy: Incidence, origin, and etiology. *Environmental and Molecular Mutagenesis, 28*(3), 167–175.

Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H. S., … Yaspo, M. L. (2000). The DNA sequence of human chromosome 21. *Nature, 405*(6784), 311–319.

Hickey, F., Hickey, E., & Summar, K. L. (2012). Medical update for children with Down syndrome for the pediatrician and family practitioner. *Advances in Pediatrics, 59*(1), 137–157.

Hirata, S., Hirai, H., Nogami, E., Morimura, N., & Udono, T. (2017). Chimpanzee down syndrome: A case study of trisomy 22 in a captive chimpanzee. *Primates; Journal of Primatology, 58*(2), 267–273.

Holubcova, Z., Blayney, M., Elder, K., & Schuh, M. (2015). Human oocytes. Error-prone chromosome-mediated spindle assembly favors chromosome segregation defects in human oocytes. *Science, 348*(6239), 1143–1147.

Hongell, K., & Airaksinen, E. (1972). A Gq deletion in a girl with Down's syndrome. *Human Heredity, 22*(1), 80–85.

Horvath, S., Garagnani, P., Bacalini, M. G., Pirazzini, C., Salvioli, S., Gentilini, D., … Franceschi, C. (2015). Accelerated epigenetic aging in Down syndrome. *Aging Cell, 14*(3), 491–495.

Hulten, M. A., Oijerstedt, L., Iwarsson, E., & Jonasson, J. (2014). Maternal germinal trisomy 21 in down syndrome. *Journal of Clinical Medicine, 3*(1), 167–175.

Ilbery, P. L., Lee, C. W., & Winn, S. M. (1961). Incomplete trisomy in a mongoloid child exhibiting minimal stigmata. *The Medical Journal of Australia, 48*(2), 182–184.

Jackson, J. F., North, E. R., 3rd, & Thomas, J. G. (1976). Clinical diagnosis of Down's syndrome. *Clinical Genetics, 9*(5), 483–487.

Jarvik, L. F., Falek, A., & Pierson, W. P. (1964). Down's syndrome (mongolism): the heritable aspects. *Psychological Bulletin, 61*, 388–398.

Jiang, J., Jing, Y., Cost, G. J., Chiang, J. C., Kolpa, H. J., Cotton, A. M., … Lawrence, J. B. (2013). Translating dosage compensation to trisomy 21. *Nature, 500*(7462), 296–300.

Karmiloff-Smith, A., Al-Janabi, T., D'Souza, H., Groet, J., Massand, E., Mok, K., … Strydom, A. (2016). The importance of understanding individual differences in Down syndrome. *F1000Research*, 5.

Kelly, S., Almy, R., Jakovic, L., & Buckner, L. (1965). Autosomal monosomy in a spontaneous abortion. *Lancet, 1*(7377), 166.

Keymolen, K., Van Berkel, K., Vorsselmans, A., Staessen, C., & Liebaers, I. (2011). Pregnancy outcome in carriers of Robertsonian translocations. *American Journal of Medical Genetics, Part A, 155a*(10), 2381–2385.

Kleschevnikov, A. M., Belichenko, P. V., Salehi, A., & Wu, C. (2012). Discoveries in down syndrome: Moving basic science to clinical care. *Progress in Brain Research, 197*, 199–221.

Korbel, J. O., Tirosh-Wagner, T., Urban, A. E., Chen, X. N., Kasowski, M., Dai, L., … Korenberg, J. R. (2009). The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies. *Proceedings of the National Academy of Sciences of the United States of America, 106*(29), 12031–12036.

Kurahashi, H., Tsutsumi, M., Nishiyama, S., Kogo, H., Inagaki, H., & Ohye, T. (2012). Molecular basis of maternal age-related increase in oocyte aneuploidy. *Congenital Anomalies, 52*(1), 8–15.

Lanfranchi, S., Toffanin, E., Zilli, S., Panzeri, B., & Vianello, R. (2014). Memory coding in individuals with Down syndrome. *Child Neuropsychology, 20*(6), 700–712.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Lejeune, J., & Aurias, A. (1976). Monozygotic heterocaryotic twinning (author's transl). *Acta Geneticae Medicae et Gemellologiae, 25*, 50–52.
Lejeune, J., Gauthier, M., & Turpin, R. (1959a). Human chromosomes in tissue cultures. *Comptes Rendus Hebdomadaires des Seances de l'Academie des Sciences, 248*(4), 602–603.
Lejeune, J., Gautier, M., & Turpin, R. (1959b). Study of somatic chromosomes from 9 mongoloid children. *Comptes Rendus Hebdomadaires des Seances de l'Academie des Sciences, 248*(11), 1721–1722.
Lejeune, J., Turpin, R., & Gautier, M. (1959a). Le mongolisme premier exemple d'aberration autosomique humaine. *Annales de Genetique, 2*, 41–49.
Lejeune, J., Turpin, R., & Gautier, M. (1959b). Mongolism; a chromosomal disease (trisomy). *Bulletin de l'Académie Nationale de Médecine, 143*(11–12), 256–265.
Lejeune, J., Lafourcade, J., Scharer, K., de, W., Salmon, C., Haines, M, & Turpin, R. (1962). Heterokaryotic monozygotism, normal twin and 21 trisomal twin. *Comptes Rendus Hebdomadaires des Seances de l'Academie des Sciences, 254*, 4404–4406.
Lejeune, J., Berger, R., Vidal, O. R., & Rethore, M. O. (1965). A case of G-G translocation in tandem. *Annales de Genetique, 8*(1), 60–62.
Lejeune, J., Rethore, M. O., de Blois, M. C., Maunoury-Burolla, C., Mir, M., Nicolle, L., … Recan, D. (1986). Metabolism of monocarbons and trisomy 21: Sensitivity to methotrexate. *Annales de Genetique, 29*(1), 16–19.
Lejeune, J. (1963). Trisomy 21 karyotypes. *Pathologie Biologie, 11*, 1153–1158.
Lejeune, J. (1966). Chromosomal studies in psychiatry. *Recent Advances in Biological Psychiatry, 9*, 13–20.
Lejeune, J. (1970). The William Allan memorial award lecture. On the nature of men. *The American Journal of Human Genetics, 22*(2), 121–128.
Lejeune, J. (1977). On the mechanism of mental deficiency in chromosomal diseases. *Hereditas, 86*(1), 9–14.
Lejeune, J. (1980). Genetics and mental welfare. In M. E. Vartanian (Ed.), *Well-being of mankind and genetics* (Vol. I, pp. 102–109). Moscow: MIR-999 Publ.
Lejeune, J. (1988). Research on pathogeny of mental retardation in trisomy 21. In *Working group on: "Aspects of the uses of genetic engineering", October 19–23, 1987. Pontificia Academia Scientiarum, Commentarii, Vol. 3, No. 31.*
Lejeune, J. (1990). Pathogenesis of mental deficiency in trisomy 21. *American Journal of Medical Genetics, 7*(Supplement), 20–30.
Lejeune-Gaymard, C. (2012). *Life is a blessing: A biography of Jérôme Lejeune*. Jérôme Lejeune Foundation (Paris, France) and The National Catholic Bioethics Center (Philadelphia, PA).
Leon, E., Zou, Y. S., & Milunsky, J. M. (2010). Mosaic Down syndrome in a patient with low-level mosaicism detected by microarray. *American Journal of Medical Genetics, Part A, 152a*(12), 3154–3156.
Letourneau, A., & Antonarakis, S. E. (2012). Genomic determinants in the phenotypic variability of Down syndrome. *Progress in Brain Research, 197*, 15–28.
Letourneau, A., Santoni, F. A., Bonilla, X., Sailani, M. R., Gonzalez, D., Kind, J, … Antonarakis, S. E. (2014). Domains of genome-wide gene expression dysregulation in Down's syndrome. *Nature, 508*(7496), 345–350.
Levitas, A. S., & Reid, C. S. (2003). An angel with Down syndrome in a sixteenth century Flemish Nativity painting. *American Journal of Medical Genetics, Part A, 116a*(4), 399–405.
Li, L. B., Chang, K. H., Wang, P. R., Hirata, R. K., Papayannopoulou, T., & Russell, D. W. (2012). Trisomy correction in Down syndrome induced pluripotent stem cells. *Cell Stem Cell, 11*(5), 615–619.
Liu, C., Belichenko, P. V., Zhang, L., Fu, D., Kleschevnikov, A. M., Baldini, A., & Yu, Y. E. (2011). Mouse models for Down syndrome-associated developmental cognitive disabilities. *Developmental Neuroscience, 33*(5), 404–413.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Lott, I. T., & Head, E. (2019). Dementia in down syndrome: Unique insights for Alzheimer disease research. *Nature Reviews Neurology, 15*(3), 135–147.

Lukowski, A. F., Milojevich, H. M., & Eales, L. (2019). Cognitive functioning in children with down syndrome: Current knowledge and future directions. *Advances in Child Development and Behavior, 56*, 257–289.

Lyle, R., Bena, F., Gagos, S., Gehrig, C., Lopez, G., Schinzel, A., … Antonarakis, S. E. (2009). Genotype-phenotype correlations in Down syndrome identified by array CGH in 30 cases of partial trisomy and partial monosomy chromosome 21. *European Journal of Human Genetics: EJHG, 17*(4), 454–466.

Macatangga, M., De la Calle, M., Torres, M. L., & Bartha, J. L. (2016). Monozygotic monochorionic twins discordant for trisomy 21: A reason to evaluate both fetuses: A case report. *Journal of Reproductive Medicine, 61*(3–4), 167–170.

Maciejewski, E., Vigneron, J., Lambert, L., Bonnet, C., & Hascoet, J. M. (2012). Down syndrome with partial trisomy of chromosome 21 because of a de-novo unbalanced translocation t(13;21)(q10;q22). *Clinical Dysmorphology, 21*(4), 200–203.

Maclean, G. A., Menne, T. F., Guo, G., Sanchez, D. J., Park, I. H., Daley, G. Q., & Orkin, S. H. (2012). Altered hematopoiesis in trisomy 21 as revealed through in vitro differentiation of isogenic human pluripotent cells. *Proceedings of the National Academy of Sciences of the United States of America, 109*(43), 17567–17572.

Martinez-Frias, M. L. (2005). The real earliest historical evidence of Down syndrome. *American Journal of Medical Genetics, Part A, 132a*(2), 231.

McClure, H. M., Belden, K. H., Pieper, W. A., & Jacobson, C. B. (1969). Autosomal trisomy in a chimpanzee: Resemblance to down's syndrome. *Science, 165*(3897), 1010–1012.

McCormick, M. K., Schinzel, A., Petersen, M. B., Stetten, G., Driscoll, D. J., Cantu, E. S., … Antonarakis, S. E. (1989). Molecular genetic approach to the characterization of the "Down syndrome region" of chromosome 21. *Genomics, 5*(2), 325–331.

Megarbane, A., Ravel, A., Mircher, C., Sturtz, F., Grattau, Y., Rethore, M. O., … Mobley, W. C. (2009). The 50th anniversary of the discovery of trisomy 21: The past, present, and future of research and treatment of down syndrome. *Genetics in Medicine, 11*(9), 611–616.

Menghini, D., Costanzo, F., & Vicari, S. (2011). Relationship between brain and cognitive processes in Down syndrome. *Behavior Genetics, 41*(3), 381–393.

Mircher, C., Briceno, L. G., Toulas, J., Conte, M., Tanguy, M. L., Cieuta-Walti, C., … Ravel, A. (2018). Growth curves for French people with Down syndrome from birth to 20 years of age. *American Journal of Medical Genetics, Part A, 176*(12), 2685–2694.

Morris, J. K., Wald, N. J., & Watt, H. C. (1999). Fetal loss in Down syndrome pregnancies. *Prenatal Diagnosis, 19*(2), 142–145.

Morris, J. K., De Vigan, C., Mutton, D. E., & Alberman, E. (2005). Risk of a Down syndrome live birth in women 45 years of age and older. *Prenatal Diagnosis, 25*(4), 275–278.

Mutton, D., Alberman, E., & Hook, E. B. (1996). Cytogenetic and epidemiological findings in down syndrome, England and Wales 1989 to 1993. National down syndrome cytogenetic register and the association of clinical cytogeneticists. *Journal of Medical Genetics, 33*(5), 387–394.

National Human Genome Research Institute. (2016). *Online education kit: Understanding the human genome project. Timeline: From Darwin and Mendel to the human genome project.*

Nelson, D. L., & Gibbs, R. A. (2004). Genetics. The critical region in trisomy 21. *Science, 306*(5696), 619–621.

Niebuhr, E. (1974). Down's syndrome. The possibility of a pathogenetic segment on chromosome no. 21. *Humangenetik, 21*(1), 99–101.

Nuffield Council on Bioethics. (2002). *Genetics and human behaviour: The ethical context–summary and recommendations*. London.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

O'Donnell, C. P., Pertile, M. D., Sheffield, L. J., & Sampson, A. (2004). Monozygotic twins with discordant karyotypes: A case report. *The Journal of Pediatrics, 145*(3), 406–408.

Papavassiliou, P., York, T. P., Gursoy, N., Hill, G., Nicely, L. V., Sundaram, U., … Jackson-Cook, C. (2009). The phenotype of persons having mosaicism for trisomy 21/Down syndrome reflects the percentage of trisomic cells present in different tissues. *American Journal of Medical Genetics, Part A, 149a*(4), 573–583.

Papavassiliou, P., Charalsawadi, C., Rafferty, K., & Jackson-Cook, C. (2015). Mosaicism for trisomy 21: A review. *American Journal of Medical Genetics, Part A, 167a*(1), 26–39.

Park, I. H., Arora, N., Huo, H., Maherali, N., Ahfeldt, T., Shimamura, A., … Daley, G. Q. (2008). Disease-specific induced pluripotent stem cells. *Cell, 134*(5), 877–886.

Parker, S. E., Mai, C. T., Canfield, M. A., Rickard, R., Wang, Y., Meyer, R. E., … Correa, A. (2010). Updated national birth prevalence estimates for selected birth defects in the United States, 2004–2006. *Birth Defects Research. Part A: Clinical and Molecular Teratology, 88*(12), 1008–1016.

Patel, J., Tan, S. L., Hartshorne, G. M., & McAinsh, A. D. (2015). Unique geometry of sister kinetochores in human oocytes during meiosis I may explain maternal age-associated increases in chromosomal abnormalities. *Biology Open, 5*(2), 178–184.

Patja, K., Pukkala, E., Sund, R., Iivanainen, M., & Kaski, M. (2006). Cancer incidence of persons with down syndrome in Finland: A population-based study. *International Journal of Cancer, 118*(7), 1769–1772.

Patterson, D., & Costa, A. C. (2005). Down syndrome and genetics - a case of linked histories. *Nature Reviews. Genetics, 6*(2), 137–147.

Patterson, D. (2009). Molecular genetic analysis of Down syndrome. *Human Genetics, 126*(1), 195–214.

Pelleri, M. C., Piovesan, A., Caracausi, M., Berardi, A. C., Vitale, L., & Strippoli, P. (2014). Integrated differential transcriptome maps of acute megakaryoblastic leukemia (AMKL) in children with or without down syndrome (DS). *BMC Medical Genomics, 7*, 63.

Pelleri, M. C., Cicchini, E., Locatelli, C., Vitale, L., Caracausi, M., Piovesan, A., … Cocchi, G. (2016). Systematic reanalysis of partial trisomy 21 cases with or without Down syndrome suggests a small region on 21q22.13 as critical to the phenotype. *Human Molecular Genetics, 25*, 2525–2538.

Pelleri, M. C., Gennari, E., Locatelli, C., Piovesan, A., Caracausi, M., Antonaros, F., … Cocchi, G. (2017). Genotype-phenotype correlation for congenital heart disease in Down syndrome through analysis of partial trisomy 21 cases. *Genomics*.

Pelleri, M. C., Cattani, C., Vitale, L., Antonaros, F., Strippoli, P., Locatelli, C., … Caracausi, M. (2018). Integrated quantitative transcriptome maps of human trisomy 21 tissues and cells. *Frontiers in Genetics, 9*, 125.

Penrose, L. S. (1954). Observations on the aetiology of mongolism. *Lancet, 267*(6837), 505–509.

Petersen, M. B., Adelsberger, P. A., Schinzel, A. A., Binkert, F., Hinkel, G. K., & Antonarakis, S. E. (1991). Down syndrome due to de novo robertsonian translocation t(14q;21q): DNA polymorphism analysis suggests that the origin of the extra 21q is maternal. *The American Journal of Human Genetics, 49*(3), 529–536.

Piovesan, A., Caracausi, M., Antonaros, F., Pelleri, M. C., & Vitale, L. (2016). GeneBase 1.1: A tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database: The Journal of Biological Databases and Curation*.

Pradhan, M., Dalal, A., Khan, F., & Agrawal, S. (2006). Fertility in men with down syndrome: A case report. *Fertility and Sterility, 86*(6), 1765.e1761-1763.

Prandini, P., Deutsch, S., Lyle, R., Gagnebin, M., Delucinge Vivier, C., Delorenzi, M., … Antonarakis, S. E. (2007). Natural gene-expression variation in Down syndrome modulates the outcome of gene-dosage imbalance. *The American Journal of Human Genetics, 81*(2), 252–263.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Rahmani, Z., Blouin, J. L., Creau-Goldberg, N., Watkins, P. C., Mattei, J. F., Poissonnier, M., … Delabar, P. M. (1989). Critical role of the D21S55 region on chromosome 21 in the pathogenesis of Down syndrome. *Proceedings of the National Academy of Sciences of the United States of America, 86*(15), 5958–5962.

Rahmani, Z., Blouin, J. L., Creau-Goldberg, N., Watkins, P. C., Mattei, J. F., Poissonnier, M., … Delabar, P. M. (1990). Down syndrome critical region around D21S55 on proximal 21q22.3. *American Journal of Medical Genetics, 7*(Supplement), 98–103.

Reeves, R. H., Delabar, J., Potier, M. C., Bhattacharyya, A., Head, E., Lemere, C., … Busciglio, J. (2019). Paving the way for therapy: The second international conference of the trisomy 21 research society. *Molecular Syndromology, 9*(6), 279–286.

Rehen, S. K., Yung, Y. C., McCreight, M. P., Kaushal, D., Yang, A. H., Almeida, B. S., … Chun, J. (2005). Constitutional aneuploidy in the normal human brain. *Journal of Neuroscience, 25*(9), 2176–2180.

Rethore, M. O. (1981). Structural variation of chromosome 21 and symptoms of Down's syndrome. In G. R. Burgio, M. Fraccaro, L. Tiepolo, & U. Wolf (Eds.), *Trisomy 21* (pp. 173–182). Berlin: Springer-Verlag.

Robinson, A. (1960). A proposed standard system of nomenclature of human mitotic chromosomes. *Current Science, 29*(9), 337–340.

Rogers, J. G., Voullaire, L., & Gold, H. (1982). Monozygotic twins discordant for trisomy 21. *American Journal of Medical Genetics, 11*(2), 143–146.

Roizen, N. J., & Patterson, D. (2003). Down's syndrome. *Lancet, 361*(9365), 1281–1289.

Roubertoux, P. L., & Kerdelhue, B. (2006). Trisomy 21: From chromosomes to mental retardation. *Behavior Genetics, 36*(3), 346–354.

Rueda, N., Florez, J., & Martinez-Cue, C. (2012). Mouse models of Down syndrome as a tool to unravel the causes of mental disabilities. *Neural Plasticity, 2012*, 584071.

Sanchez-Mut, J. V., Huertas, D., & Esteller, M. (2012). Aberrant epigenetic landscape in intellectual disability. *Progress in Brain Research, 197*, 53–71.

Schnabel, F., Smogavec, M., Funke, R., Pauli, S., Burfeind, P., & Bartels, I. (2018). Down syndrome phenotype in a boy with a mosaic microduplication of chromosome 21q22. *Molecular Cytogenetics, 11*, 62.

Schuring-Blom, G. H., Boer, K., Knegt, A. C., Verjaal, M., & Leschot, N. J. (2002). Trisomy 13 or 18 (mosaicism) in first trimester cytotrophoblast cells: False-positive results in 11 out of 51 cases. *European Journal of Obstetrics & Gynecology and Reproductive Biology, 101*(2), 161–168.

Séguin, É. (1846). *Traitement moral, hygiène et éducation des idiots et des autres enfants arrièrés ou rétardés dans leur développement, etc: J.B. Baillière.*

Shapiro, B. L. (1999). The Down syndrome critical region. *Journal of Neural Transmission, 57*(Supplementum), 41–60.

Shi, Y., Kirwan, P., Smith, J., MacLean, G., Orkin, S. H., & Livesey, F. J. (2012). A human stem cell model of early Alzheimer's disease pathology in Down syndrome. *Science Translational Medicine, 4*(124), 124ra129.

Shuttleworth, G. E. (1900). *Mentally-deficient children: Their treatment and training*. London: H.K. Lewis.

Skotko, B. G., Levine, S. P., & Goldstein, R. (2011). Self-perceptions from people with Down syndrome. *American Journal of Medical Genetics, Part A, 155a*(10), 2360–2369.

Sobol, M., Raykova, D., Cavelier, L., Khalfallah, A., Schuster, J., & Dahl, N. (2015). Methods of reprogramming to induced pluripotent stem cell associated with chromosomal integrity and delineation of a chromosome 5q candidate region for growth advantage. *Stem Cells and Development, 24*(17), 2032–2040.

Speicher, M., Antonarakis, S. E., & Motulsky, A. G. (2010). *Vogel and motulsky's human genetics: Problems and approaches*. Springer-Verlag Berlin Heidelberg.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Spellman, C., Ahmed, M. M., Dubach, D., & Gardiner, K. J. (2013). Expression of trisomic proteins in Down syndrome model systems. *Gene, 512*(2), 219–225.

Stagni, F., Giacomini, A., Emili, M., Guidi, S., & Bartesaghi, R. (2018). Neurogenesis impairment: An early developmental defect in Down syndrome. *Free Radical Biology and Medicine, 114*, 15–32.

Stahl, A., & Tourame, P. (2013). Trisomy 21 in visual art. *Archives de Pediatrie, 20*(12), 1342–1351.

Stevenson, T. N. (2009). The power of language in medicine. Case study: Mongolism. *Pharos of Alpha Omega Alpha-Honor Medical Society, 72*(4), 4–9.

Strachan, T., & Read, A. (2019). *Human molecular genetics* (5th ed.). Boca Raton, FL: CRC Press Taylor & Francis Group.

Strippoli, P., Pelleri, M. C., Caracausi, M., Vitale, L., Piovesan, A., Locatelli, C., … Cocchi, G. (2013). An integrated route to identifying new pathogenesis-based therapeutic approaches for trisomy 21 (Down Syndrome) following the thought of Jérôme Lejeune. *Science Postprint, 1*(1), e00010.

Su, M. T., Kuan, L. C., Chou, Y. Y., Tan, S. Y., Kuo, T. C., & Kuo, P. L. (2016). Partial trisomy of chromosome 21 without the Down syndrome phenotype. *Prenatal Diagnosis, 36*(5), 492–495.

Taylor, G. M., Williams, A., D'Souza, S. W., Fergusson, W. D., Donnai, D., Fennell, J., & Harris, R. (1988). The expression of CD18 is increased on Trisomy 21 (Down syndrome) lymphoblastoid cells. *Clinical and Experimental Immunology, 71*(2), 324–328.

Tolksdorf, M., & Wiedemann, H. R. (1981). Clinical aspects of Down's syndrome from infancy to adult life. In G. R. Burgio, M. Fraccaro, L. Tiepolo, & U. Wolf (Eds.), *Trisomy 21* (pp. 3–32). Berlin: Springer-Verlag.

Turpin, R., & Lejeune, J. (1961). Chromosome translocations in man. *The Lancet, 277*(7177), 616–617.

Turpin, R., Lejeune, J., Lafourcade, J., & Gautier, M. (1959). Chromosome aberrations & human diseases; multiple spinal abnormalities with 45 chromosomes. *Comptes Rendus Hebdomadaires des Seances de l'Academie des Sciences, 248*(25), 3636–3638.

Van Beukering, J. A., & Vervoorn, J. D. (1956). Monozygotic twins, one normal and the other a case of mongolism. *Nederlands tijdschrift voor geneeskunde, 100*(5), 340–341.

Van Keuren, M. L., Stewart, G. D., Bradley, C. M., Kurnit, D. M., Neve, R. L., Watkins, P. C., … Patterson, D. (1989). Characterization of an unusual and complex chromosome 21 rearrangement using somatic cell genetics and cloned DNA probes. *American Journal of Medical Genetics, 33*(3), 369–375.

Vitale, L., Casadei, R., Canaider, S., Lenzi, L., Strippoli, P., D'Addabbo, P., … Zannotti, M. (2002). Cysteine and tyrosine-rich 1 (CYYR1), a novel unpredicted gene on human chromosome 21 (21q21.2), encodes a cysteine and tyrosine-rich protein and defines a new family of highly conserved vertebrate-specific genes. *Gene, 290*(1–2), 141–151.

Vitale, L., Frabetti, F., Huntsman, S. A., Canaider, S., Casadei, R., Lenzi, L., … Strippoli, P. (2007). Sequence, "subtle" alternative splicing and expression of the CYYR1 (cysteine/tyrosine-rich 1) mRNA in human neuroendocrine tumors. *BMC Cancer, 7*, 66.

Vitale, L., Serpieri, V., Lauriola, M., Piovesan, A., Antonaros, F., Cicchini, E., … Caracausi, M. (2019). Human trisomy 21 fibroblasts rescue methotrexate toxic effect after treatment with 5-methyl-tetrahydrofolate and 5-formyl-tetrahydrofolate. *Journal of Cellular Physiology*.

Weijerman, M. E., & de Winter, J. P. (2010). Clinical practice. The care of children with Down syndrome. *European Journal of Pediatrics, 169*(12), 1445–1452.

Weisfeld-Adams, J. D., Tkachuk, A. K., Maclean, K. N., Meeks, N. L., & Scott, S. A. (2016). A de novo 2.78-Mb duplication on chromosome 21q22.11 implicates candidate genes in the partial trisomy 21 phenotype. *NPJ Genom Med, 1*.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Wilch, E. S., & Morton, C. C. (2018). Historical and clinical perspectives on chromosomal translocations. *Advances in Experimental Medicine & Biology, 1044*, 1–14.

Wong, F. C., & Lo, Y. M. (2016). Prenatal diagnosis innovation: Genome sequencing of maternal plasma. *Annual Review of Medicine, 67*, 419–432.

Yamakawa, K. (2012). Towards the understanding of Down syndrome using mouse models. *Congenital Anomalies, 52*(2), 67–71.

Yu, T., Li, Z., Jia, Z., Clapcote, S. J., Liu, C., Li, S., … Yu, Y. E. (2010). A mouse model of Down syndrome trisomic for all human chromosome 21 syntenic regions. *Human Molecular Genetics, 19*(14), 2780–2791.

Zhao, X., & Bhattacharyya, A. (2018). Human models are needed for studying human neurodevelopmental disorders. *The American Journal of Human Genetics, 103*(6), 829–857.

Zhao, W. W., Wu, M., Chen, F., Jiang, S., Su, H., Liang, J., … Yu, S. (2015). Robertsonian translocations: An overview of 872 robertsonian translocations identified in a diagnostic laboratory in China. *PLos One, 10*(5), e0122647.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari