From Genome to Genome in the Quest to Decipher Life

Gianfranco Biondi and Olga Rickards

THE GENOME

The genome of an organism consists of its entire DNA, including its genes, i.e. the material structures that contain the information for constructing proteins. Proteins, the biological building blocks of nature, determine not only an organism's anatomy and morphology but also its physiology, immune characteristics and may partly influence its behavior.

The double helical structure of DNA, or deoxyribonucleic acid, is extremely simple and repetitive. It is formed by four molecules (the bases adenine, cytosine, guanine and thymine) in a se-quence running along two opposing strands in a fixed paired arrangement: adenine with thymine and cytosine with guanine (A with T, and *vice versa*, C with G, and *vice versa*). The enormous biological variability nature achieves comes not from a vast assortment of molecules but derives instead from the particular linear order in which the one half of the four base pairs is arranged along a strand. This in turn determines the order in which they join the complementary base pair members on the other strand. The genomes of living organisms are composed of two fragments—coding and non coding segments. In the coding segment that specifies proteins, each sequence of three bases determines which amino acid will be located at a certain position in a certain polypeptide chain. The non coding segment, which was long thought to be useless and hence dubbed "junk DNA", consists mainly of repeated intergenic sequences and introns and pseudogenes, i.e. non expressed genetic DNA. Although the function of this part of the genome, called "the other genome", is not yet fully clear, researchers have begun to understand the crucial role and diverse functions it may have in keeping coding DNA working properly and in regulating cell development and differentiation. What we do know much more about is the random event that during the course of evolution has so often modified the conservative structure of an organism's DNA, making some organisms human beings and others butterflies, and yet others roses or grain—mutation. Mutation is the underlying mechanism by which one base is substituted for another. If substitution occurs in a coding region of DNA, then its linear order is changed and likewise the protein that depends on a particular gene.

All living organisms are related through the unique structure of DNA and the specific common sequences they share. So the genetic code is the same for viruses as it is for bacteria, plants and animals, including humans. However, each species has its own genome, which can also become the object of a "genome project". The design of such a project is not tied so closely or only to our curiosity about the world around us as it is to the knowledge acquired from studying non human genomes and then applied to clarify certain aspects of human biology and the ways in which it works. A metaphor for the biotechnological research supporting the Human Genome Project could be represented by the figure of Justice holding up the scales of justice in one hand. The one end of the scales would be weighed down by narcissism and utilitarianism, tipping up the other end that holds curiosity about our place in nature.

Contrary to common assumption, DNA does not self replicate, so it can be considered a "dead" molecule with few chemical properties. But thanks to this limitation, it can be collected and sequenced from blood stains and other dried organic fluids for forensic analysis or from ancient tissues (bones, teeth, hair, mummified tissue) dating up to ca. 100,000 years ago for phylogenetic studies (even though, DNA should potentially survive up to 1 million years). Rather than producing proteins, DNA is the product of a complex cellular mechanism involving particular proteins, the enzymes. By itself, DNA is nothing but an inert template from which other copies are made. Organic molecules cannot self reproduce. For it is only whole cells that possess all the structures necessary for "self" reproducing the genome. DNA can make copies of itself only if assisted by other cell mechanisms, and only after its double helix has been split apart.

So while DNA is unable to do anything by itself, many interesting things can be done with DNA. The linear sequence of its bases is used by the cell to define the order of the amino acids making up the proteins and to decide when and where they should be synthesized. But these proteins are, in turn, made up of other proteins in a seemingly never ending trail. This impression derives from the simplistic idea that only genes are passed from parent to offspring.

Actually, the egg is a complex set of machinery for reproduction ready to start even before it is fertilized. So what we inherit is not only our parent's DNA but also the entire machinery that goes along with it. Since the double helix is merely a string of information read by cell operators during the production process, only a deformed ideology could transform it into an engineering project, plan or design for construction.

HISTORICAL BACKGROUND

The story began during the 17th century when Luiz de Mercado writing in *De morbis hereditariis*, recognized that both parents and not only the father, as Aristotle thought, contributed to forming the seed that generated offspring. At the time, Marcello Malpighi suggested that the egg contained a completely preformed living organism. On the basis of work by Johan Ham on sperm cells, Anthony von Leeuwenhoek and Niklaas Hartsoeker turned Malpighi's hypothesis around to claim that the preformed organism was contained in the father's sperm instead and was only nourished by the mother. The debate continued until the middle of the 18th century, when Caspar Friedrich Wolff staked an authoritative claim for empirical research in this area of the natural sciences. Founded on the firmer ground broken by Wolff, the fundamental studies on plant heredity conducted by Carl Friedrich von Gaertner and Joseph Koelreuter opened the way for Gregor Mendel's work.

Around the early second half of the 19th century, the founder of eugenics and genetic biometrics, Francis Galton claimed in his essay *Hereditary Talent and Character* published in *Macmillan's Magazine* in 1865 that not only physical characteristics but also mental traits could be inherited. But the crucial turning point came with Mendel's presentation of his work on plant hybridization before the meeting of the Naturforschender Verein (Association of Natural Sciences) in Brno on February 8th and March 8th , 1865. His presentations were published a year later in essay form in the Acts of the Proceedings. Mendel discovered the discrete units that allow the transmission of characteristics from one generation to the next and the laws governing heredity. These hereditary factors were later termed "genes".

Building on Mendel's work, many other researchers contributed to the understanding of the mechanisms underlying inheritance. In 1875, Oscar Hertwig in his studies on animal breeding found that there was a continuity of nucleus cells which he defined as *omnis nucleus e nucleo*. Between 1880 and 1882, Walter Flemming discovered that sister chromatids separate during cell mitosis. Eduard van Beneden in 1883 demonstrated that chromosomes distribute equally in the nuclei of daughter cells. In 1888, Theodor Boveri proved that each chromosome pair represents a separate entity, and Wilhelm von Waldeyer-Hartz coined the term "chromosome". Towards the end of the century, Karl Nägeli formulated the concept of "idioplasma", i.e. the idea that a small amount of "plasma" contained the information needed to generate offspring, but he failed to find which part of the cell was responsible for storing this information. August Weismann found that during meiotic division chromosomes duplicate according to an exact model. By the beginning of the first decade of the 20th century, the knowledge basis had been set for Thomas Hunt Morgan to conduct studies on the drosophila, the common fruit fly, and to find that the genes were located on the chromosomes; only later was he able to demonstrate that the genes were arranged linearly and that their position could be perfectly identified and mapped.

Between the 1920s and 1940s, studies showed that X rays could induce genetic mutations and that the genes consisted of deoxyribonucleic acid. It wasn't until 1953 that James Watson and Francis Crick were able to describe the molecular structure of DNA—two nucleotide strands, each made up of a sugar and a phosphate group and a base, twisted in a spiral that formed a double helix. Towards the end of the 1950s, Sydney Brenner, Matthew Meselson and Francois Jacob discovered messenger RNA, the molecule that copies information from DNA to construct a polypeptide chain and that transfers it to the "cytoplasmic factories" where the chain is actually assembled. These "factories" are the organelles known as ribosomes. Ten years later, researchers identified a particular class of enzymes able to recognize a short sequence of base pairs and to cut DNA at these sites. Called restriction enzymes, in 1973 they were used to separate a fragment of animal DNA that was then transferred to a bacterium, where it was copied over and over again. This technique of clonation allowed researchers to study genes in greater detail. Four years later, Allan Maxam, Walter Gilbert and Frederick Sanger independently developed methods for sequencing DNA, i.e. processes for reading the succession of the four chemical bases that alternate along a strand of the double helix. In the early 1980s Kary Mullis developed one of the most extraordinary biotechnology methods—polymerase chain reaction (PCR)—

which permitted the rapid generation of millions of copies of a DNA molecule in just a matter of hours. For this discovery he was awarded the Nobel Prize in 1993.

PRODROME OF THE HUMAN GENOME PROJECT

A proposal to map the entire human genome was first advanced in an article by David Botstein, Ronald Davis, Mark Skolnick and Ray White that appeared in the *American Journal of Human Genetics* in 1980. At the time, although the technique for sequencing DNA was known, it was not yet fully developed and was employed only in research laboratories. So the suggested approach was to map the genome using restriction fragment length polymorphism (RFLP). To do this, many restriction enzymes had to be used. Because each enzyme recognizes a well known and short sequence of base pairs, the restriction enzymes could cut DNA into fragments of various length (hence the name RFLP) to produce a map of the base sequence of the DNA being studied. During the 1980s, restriction enzymes become the workhorses of genome research and today they are still used, together with DNA sequencing studies.

Starting in 1985, the quest for discovering all the bases making up the human genome became a kind of obsession for many molecular biologists. Charles De Lisi, co-director for medical and environmental research at the US Department of Energy (DoE) opened the discussion to create what would turn into the most ambitious and costly biological research project ever undertaken. The DoE began to dedicate funds to the set up of the project in 1987—US$5.3 million, with an estimated budget of US$1 billion for the next seven years. In the meantime, Nobel Prize laureate Renato Dulbecco was promoting the program on the pages of *Science* and at a conference titled *The Molecular Biology of Homo sapiens* held at the Cold Spring Harbor Laboratory in New York State. The emerging prospects for profits and that perhaps the biggest deal of the next century was about to take place incited the private business sector to begin investing in biotechnologies. In 1982, Akiyoshi Wada speculated about the possibility of automatically sequencing DNA and approached Hitachi with a plan for the construction of robots that could do this. In 1985, Kary Mullis developed the PCR method at Cetus Corporation, and the following year Leroy Hood and Lloyd Smith, while working at the California Institute of Technology (Caltech), invented a prototype system for automatic sequencing. In biology as in chemistry and physics, the time was ripe for the transformation of scientists into entrepreneurs. Caught between the contending interests of science and business, it mattered little whether research was devoted to gene sequencing, knowledge of which belongs to all of us, or to successfully sending voice or images over distances. In 1987, Walter Gilbert quit his job at the US National Research Council (NRC) to set up his own company, the Genome Corporation, in order to be able to patent the results from research into human genome sequencing and then sell them for profit. That same year the first automatic sequencer (Applied Biosystems) appeared on the market.

The proposal to launch the study of the human genome was beginning to take shape. In 1988, the NRC and the Office of Technology Assessment (OTA) estimated an annual budget of ca. US$200 million for the project. Soon thereafter, the National Institutes of Health (NIH) decided to take on a guiding role in the project and created the Office of Human Genome Research, which was transformed a year later into an institute wielding more prestige and power—the National Center for Human Genome Research (NCHGR). James Watson, who with Francis Crick had discovered the double helical structure of DNA, was appointed to direct the institute. In his address before the US Congress, Watson stressed the importance of allocating 3% of the project budget for studies into the ethical, legal and social aspects that knowledge of the genome would raise. In this way, he managed to secure political support for the project. Two other important events marked that year. The First Annual Genome Meeting, a kind of blessing bestowed by the academic community, was held at the Cold Spring Harbor Laboratory, and in October the NIH and the DoE signed a joint protocol of intention that opened the way for scientific cooperation between the two bodies in research on the human genome. This cooperation was consolidated a year later with the establishment of a joint committee whose task was to define in what ways the emerging consequences —the ethical, legal and social questions Watson had raised— needed to be faced.

THE PROJECT

The Human Genome Project (HGP) formally started on October 1st ,1990 with the merger of two biotechnology research programs the DoE and the NIH had been pursuing. Seven years later, the DoE

created the Joint Genome Institute and the NCHGR was transformed into the NHGRI, so that when we speak of the HGP, this refers to the combination of the two programs known as the DoE Human Genome Program and the NIH Human Genome Research Institute (NHGRI). Research into genetics and molecular biology had always been part of the institutional tasks of the NIH, which are responsible for the surveillance and improvement of the health of the US population. A less explicit part was that of the DoE, whose distinct interest lays in the evaluation of long-term effects of radiation and chemical agents on human health.

The original planning of the HGP included a total investment of US$3 billion and 15 years of work to determine the sequence of just over 3 billion base pairs that constitute human DNA —subdivided into 23 chromosomes— and to identify the genes which at the time were estimated to be about 100,000 but today are known to number no more than 35,000—40,000 (the same as the puffer fish, *Fugu rubripres*, and roughly double that of the fruit fly, *Drosophilia malanogaster*, and the nematode worm, *Caenorhabditis elegans*). The rapid technological progress that marked biological research, however, held promise that the program would be completed by 2003, exactly 50 years after the discovery of the structure of DNA. But already on June 26th 2000 a "working draft" of the genome sequence was publicly announced and simultaneously published in *Nature* and *Science* in February 2001. The framework of the HGP also comprised the study of the DNA of a group of non human organisms (e.g. the species *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*) which were to serve as a comparative model to better understand how the human genome works.

Following a conflict that arose in 1991, Watson left the NCHGR in 1992, and Craig Venter quit his job at the NIH to found the Institute for Genomic Research (TIGR), a private non profit biotechnology company whose products would be sold to the associated company Human Genome Sciences. In this way, the private sector gained a foothold in the study of the human genome, a prime example of which was Celera Genomics Corporation founded in May 1998. Venter took this occasion to declare that his company would sequence the entire human genome with an investment of US$300 million over the next three years. In the meantime, the US public consortium had acquired an international scope, enlisting the efforts of laboratories in Great Britain, France, Germany, China and Japan.

THE GENOME MAP

DNA is a highly repetitive molecule in about 50% of its base pairs, so that various regions of the human genome are over 98% identical, although physically distant from one another on the same chromosome or even located on different chromosomes. For this reason, scientists thought a genetic map, i.e. a physical map of the genome, needed to be created before they could start to analyze the DNA sequence. In fact, it was impossible to sequence the entire DNA in one go, if for no other reason than DNA is fragmented in various chromosomes. And not even that of a single chromosome could be sequenced in its entirety. According to John McPherson, co-director of the Genome Sequencing Center of Washington University, St. Louis, Missouri, it was like having a picture puzzle of a forest scene with many trees that looked alike. This meant that it would be easier to first assemble the pieces of each tree and then fit the trees together. To properly study the genome, it had to be divided into many pieces, then each piece examined separately before they could be assembled into a string. But the risk of working without a map—a risk that had to be absolutely avoided—was that the location of similar but not identical base lengths might be misarranged, with the result that strings would be created that did not perfectly fit the topography of the true human genome, perhaps with pieces of various genes that could form chimera. The map provided the basis for choosing the portions of DNA to be sequenced and the grid on which they would be assembled and the exact location determined where each one fit the others along a chromosome.

The genome map was created using bacterial artificial chromosomes (BAC), i.e. DNA segments about 175,000 base pairs long that were mass cloned. Each DNA segment had to be cloned in order to have all the copies necessary so that the researchers would be able to arrange the various types of clones into a series with numerous overlapping points that was long enough to cover most of the genome. This was achieved to 95%, thus covering about 96% of the entire genome. Each type of clone was treated with various restriction enzymes and the resulting configurations were then recorded. Each restriction enzyme recognizes a short, specific sequence of bases where it cuts the DNA and fragments the clone into two or more segments of different length and weight that can be easily identified.

The collection of all the pieces produced by a battery of enzymes provides what can be called the clone's "fingerprint", an identification mark that distinguishes it from other clones and can also show which clones contain the same sequences to be overlapped. In this way, milestones were set along the DNA to mark the short base sequences—the map. In order to make the map as detailed as possible, huge amounts of clones were created and studied, each with its own specificity, so that they could be overlapped from one point to the next, so covering the entire genome about 20 times over. The most representative clones were then sequenced to determine the exact order of the copies of the base pairs, and then assembled by computer programs, each one according to its location on the map. Once assembled, the chain of overlapping or contig clones (contiguous sequence of DNA created from the assembly of contiguous chromosome fragments, both natural and artificial, like the BAC) defined the string of the human genome.

EXPLORATION OF THE GENOME

The distribution of genes within the genome does not follow a uniform model in all living creatures. In the chromosomes of mammals, regions of high genetic density alternate with those where single genes lie far apart from one another, and those where only non-coding DNA may be found, the "other genome".

The metaphor suggested by Robert Waterston, director of the Genome Sequencing Center at Washington University, St. Louis, to describe the genome's topography or the "genomic landscape" refers to an urban environment, with a densely populated center surrounded by lesser and lesser developed areas spreading outwards, passing from isolated farms dotting the prairies to desert expanses. In other organisms, the genome is organized like the outskirts of big cities—compact yet irregularly dislocated.

The "metropolitan" areas of the human genome are rich in cytosine and guanine. Under the microscope, they appear as pale bands. The C-G pair is mostly concentrated in the DNA segments that contain up to 30,000 copies of the two pairs; they are highly repetitive, are located next to genes and work as regulators. In contrast, the A-T pairs are more present in the "desert" areas that appear as dark bands. Another particularity regards the clusters of HOX genes, homeotic genes expressed along the antero-posterior axis of the embryo, which play a fundamental role in the regulation of growth but do not contain the "other genome". This may be attributable to the effect of some unknown evolutionary cause that was able to secure the homogeneity of these gene clusters.

Less than 2% of the human genome codes for proteins, whereas the bulk of it —at least 50%— consists of repeated sequences of the "other genome" which have preserved the memory of genetic evolution. As such, they constitute true storehouses of "fossil" genetic information. The repeated sequences derive from transferable elements —like those of viral genomes— that proliferated by random insertion along the DNA structure. They restructured the genome by creating new genes or modifying and reshuffling the existing ones. By dating them, the time they were formed can be determined, the characteristic genealogical relationships can be established and so the course of our hereditary history defined. Within the human genome, a huge amount of repeated sequences have accumulated which are much older than those observed in other living creatures. This circumstance could depend on the inefficiency of our system to clean up our DNA, or more exactly, the inefficiency that seems to characterize mammals. In fact, the half life of some of these structures indicates that the last cleaning up took place more recently in insects than in mammals. Over the last tens of millions of years, the rate of accumulation of repeated sequences in the evolutionary line that led to the human species has slowed down considerably. A reduced pace has not been observed for rodents, however. Parts of our oldest "other genome" have become extinct while others are on the brink of extinction. But both —the DNA transposones and the RNA transposones, i.e. the Long Terminal Repeats (LTR)— have survived in the genome of mice, suggesting that this behavior might be linked to certain differences between the two groups of organisms. What might have been at the origin of the phenomenon was population dynamics and structure. Rodents have always lived in large communities, unlike our direct predecessors who must have faced frequent bottlenecks at certain moments in their history. Consanguinity and genetic drift might have played a role here as well.

The repeated sequences are located in the genome regions rich in A—T and poor in C—G. This rule, however, is subject to exceptions as concerns a particular family of repeated DNA, the SINE sequences (Short Interspersed Elements), and particularly the Alu, which present cutting sites for the Alu I enzyme. They have found their place where

genes cluster together, i.e. where the bases cytosine and guanine are highly concentrated. There are two possible explanation for this. We can speculate that in some way the SINE sequences "discovered" how to move around, or that at the beginning they developed in the "right" position (where adenine and thymine are abundant), but later evolution could have promoted those sequences that had wandered to the other part. Evidence for the former hypothesis is the observation that the more recent Alu sequences are located exactly where the A—T pair is abundant, whereas the older ones appear in regions with more C—G. Hence, also the latter hypothesis conferring a selective advantage on these sequences in the coding sectors seems founded.

Thanks to the dating and analysis of the model of dispersion of repeated sequences it was possible to show that there is a rate of differential mutation between the two sexes of our species, with a substitution rate in the male germinal line that is twice that in the female line. Among the many reasons for this difference may be the greater number of cellular divisions involved in the formation of spermatozoa compared with that of eggs and the different repair mechanisms that distinguish the two types of germinal cells.

The lower number of genes present in our genome —only one third of the total estimated five years ago, and just over twice that of an insect— raised the problem of how to account for the enormous complexity a species presents despite such a modest inheritance. The answer is that each human gene is able to contribute to the production of more than one protein. For humans, the rule of "one gene per polypeptide chain" is replaced by that of "one gene for an average of three different polypeptide chains". Through a mechanism of alternative splicing, the products of an exon of the same gene can combine in various ways and thus provide variants of a protein starting from the base components. This is possible because our genes are scattered along the entire DNA structure and the coding regions (the exons) for the various parts of the same protein are not necessarily near one another. This is how a gene —or its "pieces" to be more exact— manages to build three proteins. This type of efficiency is unknown to the insect world and to other living creatures as well.

The ability to rearrange existing proteins in order to build others that are better at new functions was the evolutionary process that enabled humans and vertebrates in general to have a proteoma —the collection of proteins produced by a genome— that was complex yet kept a relatively simple genetic structure. Over half of our protein families can be considered true superfamilies, each with a high number of family groups whose origin can be traced back to genetic duplication, probably an important evolutionary force in the history of vertebrates. Two significant examples of this are immunoglobulin (Ig) families, which are so versatile that they can mount a wide spectrum of immune defenses against infection, and epithelial proteins (e.g. keratin) which strengthen and support body organs. However, the human genome has also undergone the opposite process—contraction of genetic families. An obvious example of this is our sense of smell, whose receptors belong to an important family with over 1000 members. Unlike most other vertebrates, our survival depends relatively little on having a highly developed sense of smell, so that half of the human olfactory receptors have lost their function since the genetic information the receptors received has become silent.

About 200 genes have been found to have analogues in bacteria but not in invertebrates. Evidently, they integrated into the chromosomes of our evolutionary line following bacterial infection after vertebrates and invertebrates had developed. Transferred genetic material serves important physiologic functions ranging from brain activity to immune response, and thus must have represented a selective advantage for early vertebrates since they kept the material inside the genome down to humans. Known as horizontal transfer, this process today seems highly unlikely because our germinal cells (eggs and spermatozoa) are isolated from the external environment and humans have developed a sophisticated immune defense system against extraneous organisms. But in the early phases of the evolutionary process, defense against parasitic invasion —a form of primordial migration— was probably weak and so bacteria were able to enter higher level hosts and exchange genetic material with them, perhaps through viral mediation.

Compared with other living creatures whose genome has been sequenced, humans have many more proteins involved in cell structure, immune defense mechanisms, DNA copying, synthesis of RNA and proteins, and intracellular communication.

Genetically, human beings differ from one another only by about 0.1%, which means one base every 1000; this difference takes both the peculiarity of each individual into account and the genetic basis of diseases. Within the framework of the HGP, the members of the International SNP Map Work Group, which involves five research centers, have discovered

1.4 million single bases that present with mutations—only a fraction of the total—and defined their exact position in the genome. Each single point of variation constitutes a Single Nucleotide Polymorphism (SNP). These polymorphisms are unevenly distributed; certain DNA segments have many, while others have none. Why this is so may be explained in evolutionary terms, in that a lack of polymorphism could mean that certain functions were essential, stable genetic forms, i.e. those selected over the course of time, and if altered could have been incompatible with life. One example of low variability is chromosome X, whereas the HLA region, which codes for proteins of immune response, is highly differentiated. Over 80% of SNP occurs in over 10% of the population. Having a map of them is useful for the study of genetic diseases. But anthropology as well will undergo a major revolution, because our evolutionary history can now be reconstructed in greater detail and with greater precision. The first results have already arrived with the confirmation of a model that postulates the origin of humanity as starting from a small population (ca. 10,000 individuals) that grew rapidly and populated the earth about 100,000 years ago.

THE "OTHER GENOME"

The first surprise the HGP held in store for us was the discovery that we possess so few genes and that our genome is about as large as that of many other animals, including our phylogenetically distant relations. This demonstrated that human beings and other living creatures differ not because of the qualitative differences between genes but rather by how the genes aggregate and are regulated, activated and disactivated during the course of their development and cell differentiation according to a complex scheme.

The second surprise was the "other genome", which was thought to be inert and so was called "junk" DNA. Actually, we are starting to understand that DNA has a meaning in the composition, organization and function of the genome.

About 75% of "other genome" is made up of intergenic DNA, of which 50% is repetitive DNA and 25% is non expressed DNA, i.e. introns and pseudogenes. The close association between high density gene regions and high density of Alu sequences support the hypothesis that Alu sequences play a role in the regulation of gene expression. Similarly, the high number of repeated sequences in the regions that control chromosome function, like the centromeres and the telomeres, seems to strengthen the idea that they are involved in securing the proper segregation of chromosomes during cell division at mitosis and meiosis.

Another important consideration that sheds light on the role played by the "other genome" is the quantity of DNA that makes up a genome, in other words its size (gs = genome size). Gs values vary greatly among plants and animals, basically in relation to the level of non coding DNA present in each species and in particular to mutation of the number of repeated sequences. The changes we observe, however, do not seem to be very closely correlated to the anatomo-functional complexity of living organisms. In fact, if among prokaryotes the *Bacillus megaterium* has a genome of 30 million bases, in fungi the *Saccharomyces cerevisiae* has only 12 million, in invertebrates the *Drosophilia melanogaster* has 180 million, while the *Locusta migratoria* has 5 billion, and among vertebrates, humans have a genome of 3 billion bases and the *Necturus maculosus* 160 billion. As gs increases, so does non coding DNA, whereas gene density decreases, with 483 genes per million bases in *Saccharomyces cerevisiase*, 117 in *Drosophilia melanogaster* and only 15 in *Homo sapiens*. The comparison between species now proposed leads to a suggested role of the non coding fraction of DNA in the functional genome, i.e. in all that concerns the architectural and regulative tasks the genome has to do.

The "other genome" also seems to have nucleo-typical and regulative effects. The former can be identified by comparing the gs of the bat *Myotis capaccinii* with that of the rat *Proechimys guairae* —1.9 *vs.* 6.3 picograms, respectively. In fact, the gs is correlated with the volume of the nucleus and that of the cell. Clearly, a small genome is more adapted to the metabolic needs of flight, where gas exchange is more efficient due to the greater surface area provided by minute cells. The same relation is also found in birds, so that flying animals have smaller genomes than those of non-flying animals. The best example for nucelo-typical action is observed in the plant species of warm-hot climates, in which low gs values shorten germination time. The second effect is linked to the high number of repetitive sequences present in a genome.

In bats and puffer fish, the non-coding fraction, i.e. that of repeated sequences, barely reaches 10% of the total, indicating that the genes for some complex physiologic functions most likely need less of this DNA to express normally during

development. Repetitive DNA, therefore, seems to have an important effect on determining the phenotype though not on the ability to code for proteins. We are still far from having conclusively identified the meaning and role of the "other genome", but the HGP has shown that some genetic functions are regulated by non-coding DNA.

THE GENOME AND ITS UTOPIA

Living organisms are made of material. But the organization of this material is unique. Most biologists believe that the primary goal of research is to try to understand what life is and see the manipulation of living matter as a secondary objective. However, other scientists with a different research paradigm, especially those that create the so-called "public opinion", have made intervention in the biological world the center of scientific enterprise. Ultimately, this represents the continuous tension between two characteristically human impulses: that of understanding and that of acting.

One of the major ambitions of the HGP, or more precisely its utopia, was to think it could completely define what its means to be human once the workings of our genome had been revealed. For those who believe that DNA creates both mind and body, genetic variation and combinations of genes determine the vast variety observed among individual members of our species, i.e. an explanation for why some people are successful, while others fail in society, enjoy good health or fall ill. In brief, the HGP is determined to reveal how life works and to gain a philosophical grasp of ourselves. But a living creature is more that its DNA; it is the unique result of an evolutionary process that involves genetic and environmental information. Environmental information is partly the consequence of an organism's activity, which produces and consumes the conditions of its existence. The world of plants and animals is not pre-formed the moment they are born. They make it up.

The tension between studying nature and controlling it, two aims of biological research, also resides within the HGP. In fact, once the simplest biological problems have been resolved, e.g. taxonomy, the cost of taking on the more complicated ones rises considerably; it can be difficult, or at least so politicians believe, to justify enormous investments for merely cultural purposes. This is why the focus of interest in sequencing our genome has been shifted to the possibility of treating human diseases. This other important objective of the HGP, namely to obtain information from the complete sequencing of our DNA, would provide the basis to enable 21st century medicine to intervene in the over 5000 genetic diseases that affect humanity and those in which hereditary factors play a major role besides. Only a centrally coordinated project with specific objectives was thought to be able to furnish the indispensable guarantees to achieve this goal. On one point the program placed particular emphasis—and since it was a public program, the emphasis was truly significant and sounded like a guarantee. The research results would be entered into electronic databases in order to make the information easily accessible and in the most suited form to anyone who was interested in using it. These criteria, first formulated in February 1996 at a conference the British Wellcome Trust held in Bermuda, have become known as the "Principles of Bermuda". Interestingly, they were not accepted by the private sector.

THE WORLD ANNOUNCEMENT

Speaking from the White House on June 26th 2000, President Bill Clinton, together with Tony Blair, announced that the human genome had been completely sequenced by the public consortium, the HGP, and the private company, Celera Genomics Corporation. Clinton defined the results the scientists had obtained as a historic event, as it would lead to a new era in molecular medicine. An era in which new strategies for prevention, diagnosis and treatment of disease would open henceforth. He underlined the need to guarantee the confidentiality of genetic information through a safeguard system that would protect individuals or groups of individuals against discrimination.

The entire announcement focused on health care issues. Heading the President's prognosis was the claim that knowledge of the genome would indicate to scientists which genetic variations caused a certain disease, so that healthy persons could undergo clinical examinations to determine whether they were at risk for developing the disease. Then he stated that it would be possible to predict the course of disease and to establish more precise diagnoses, an element absolutely indispensable for securing the most suitable treatments and the most effective drugs in order to personalize treatment as far as possible and reduce side effects. Lastly, he explained that the identification of the molecular causes of disease would render therapeutic interventions more effective, since it would be possible to intervene by replacing the defective gene with a healthy one, or more simply, replace the defective

protein product with a normal one produced synthetically. It should be remembered that, while acclaiming the new season of molecular medicine opened by the HGP, the President did not forget to approve the Genetic Nondiscrimination in Health Insurance and Employment Act of 1999.

Several passages of Clinton's speech explicitly showed how deeply rooted irrational thought is in our otherwise so scientifically and technologically advanced society. In reference to Galileo, who, by mathematics and mechanics described the motion of celestial bodies and explicated the language used by God to create the universe, Clinton extended this idea to the HGP, declaring to all the world that, "Today, we are learning the language in which God created life. We remain all the more impressed by the complexity, beauty and wonder of this divine and sacred gift." No more eloquent a eulogy for evolution had even been spoken before eminent biologists. But we should not be surprised when even Walter Gilbert entitled his article in the book *The Code of Codes: Scientific and Social Issues in the Human Genome Project* (eds. D.J. Kevles and L. Hood), "A vision of the Holy Grail". And the volume editors repeated the metaphor, advocating in the preface that "The search for the biological grail has gone forward, starting from the beginnings of the century, and has now entered its culminate phase with the recent creation of the human genome project".

While certain socio-political issues addressed in the President's announcement lend it weight and brilliance, much of the text is perfused with spiritualistic cant. For example, "The horizon that lies before us is one that science cannot face alone. It is the horizon that represents the ethical, moral and *spiritual* (our italics) dimension of power that we now possess. We cannot avoid exploring this outermost frontier of science. But while reflecting on how to use these new discoveries, we cannot be held back by our oldest and most cultured human values. We must ensure that the new science of the genome and its benefits are aimed at improving the life of all citizens of the world, and never a privilege for the few […]. The growing knowledge of the human genome should never change the basic principles on which our ethics, our government, our society are founded. We all are *created* (our italics) equal and bear the right to be treated equally before the law. After all, I believe that one of the major truths to emerge from this triumphant exploration of the genome is that in genetic terms, all human beings, regardless of *race* (our italics to show that it is incorrect to speak of human races) are 99.9% equal".

Blair, not wanting to take a back seat, stated, "All of us share the duty to ensure that the common property of the human genome is used freely for the common benefit of the entire human *race* (our italics to underline the fact that there is no *human race* but rather a human species), to ensure that the power of information now at our disposal will be used to transform medicine and not abused to make man a *creator* (our italics) of his own kind or to invade the individual's right to privacy".

The results of research on the human genome were published separately by the public and private consortia. On February 15th 2001, the HGP published the sequence in the journal *Nature* and the next day Celera Genomics Corporation published their version in *Science*.

A PROJECT FOR HUMAN DIVERSITY

Many scientists have pointed to the rather considerable limitation to HGP of not having considered genetic variation —the characteristics that make life possible in the first place. Starting from this consideration, and to fill this gap, in 1991 Luca Cavalli Sforza and other researchers advanced the proposal to create within the HGP a section to study genetic variation in individuals from various ethnic backgrounds. The program, called the Human Genome Diversity Project (HGDP), had the double aim to conduct research on human origins and evolution and to work as a base for the planning of research into complex diseases. After a lengthy series of controversies lasting over five years, in October 1997 the project was approved by the US NRC, although encumbered by several suggestions that limited its portent. According to the examination committee, the international character of the HGDP had to be reduced and its scientific ambitions curtailed as well. In particular, the plans concerning the research into new genes involved in disease were to be scotched because this type of objective would complicate the logistics of the project and make it difficult to protect the anonymity and other rights of DNA donors. In the opinion of the committee president, William Schull of the School of Health, University of Texas, Houston, the project deserved support, although the ethical and legal problems that it raised could hinder its realization.

The HGDP proponents stated their moderate satisfaction with the partial recognition, as their main but not sole interest lay in studying human evolution, and they were perfectly aware of the need to safeguard the rights of the individuals and the communities

studied. But despite this statement, which gave the appearance of permission to go forward, until 1999 no funds were made available either by the National Science Foundation (NSF) or the NIH. Furthermore, the NIH advanced the conjecture that there might be other approaches to study genome variations; with this tack they resized the HGDP. It is very sad to note that both private and public institutions alike are reluctant to invest money in cultural projects such as the study of the reconstruction of the evolutionary phases of our species. Apparently, what is most important to the Western nations is to win the war on cancer and other diseases. The fight against disease must continue, and we too share in this effort; even so, there are also other causes to be advanced. We are discontent with the extremely utilitarian role biology has taken upon itself. We believe that to study our place in nature is at least as valuable as it is to do medical research—and to be truthful, it means more to us.

In the report, *Evaluating Human Genetic Diversity*, drawn up by the NRC, they underlined the need to reduce the ethical and legal complications to a minimum by selecting only unrelated donors and by adopting a cataloging system that made it impossible to associate donor and donor sample. A similar procedure showed the additional advantage of relieving scientists of having to obtain consent for each future study. But this strategy has made the data completely useless for biomedical purposes, because in order to discover the genes associated with a disease, it is also necessary to know the state of health of the individual and to analyze the DNA of the greatest number of family members carrying a disease. In fact, there is a higher probability of finding many more that share the same inherited defective trait among family members than in the rest of the population and, therefore, the gene responsible for the disorder. However, because it is the discovery of "bad" genes that drives the development of a cure with potential market value, with the consequent risk of speculation by the pharmaceutical industry, anonymity is the only guarantee against ethical and legal entanglements. In brief, we have a serpent that bites its own tail — anonymity affords social advantages yet sets severe biomedical limitations.

The HGDP was designed to study the role of various aspects that characterize the structure of population, models of migration and reproductive habits, and to determine which genetic mutations are associated with a particular disease. To reach this goal, DNA samples are needed —blood, hair and cells of the oral mucosa—from hundreds of subjects belonging to about 500 populations around the world and from several families. To show the degree of genetic affinity between populations, the best strategy would be to analyze the same genome segments of each individual. But it is expressly on these points, i.e. which populations and which genes to choose, that no agreement has yet been reached. A further question concerns the way to involve indigenous groups in the benefits, also the economic benefits, that could accrue from the study of their DNA and that can transform the findings into medicines.

Another thorny problem is that of complex diseases in which several genes interact with the environment to determine the pathological syndrome. In the attempt to single out the various defective components, the strategy has been to work on small, isolated communities where the symptoms of particular pathologies are very common. The first example is the study of diabetes in a linguistically isolated population living in western Finland, which was assured participation in the economic and health benefits, in which the same segment of chromosome 12, about 20 million base pairs long, was found in many related subjects with diabetes. As in other instances of complex diseases, the most difficult part of research is to determine in such long genomic regions the exact gene or genes responsible for the disorder. Perhaps the most promising possibility could come from the discovery of which genetic markers or DNA segments are associated with which illness, i.e. to evaluate the genetic imbalance. This is none other than the degree of association of specific sequences in a population. For our purposes, the association depends on the distance between marker and gene, marker polymorphism, its mutation rate and level of recombination, which is a function of the distance that separates the two elements in a specific region of the genome. Some of these parameters are also directly influenced by the distinctive structure of a population, its size and degree of endogamy.

Obviously, unlike what the HGDP had planned, the study could be performed on only one population at a time. But much richer information can be obtained and greater universal validity achieved only if the organization of the imbalance is seen together with the largest possible part of the genome and in populations with different histories. This ploy was skillfully used by the pharmaceutical company Sequana Therapeutics of La Jolla, California, to discover the genes responsible for asthma after it

had funded the collection of information of various small populations from the Tristan Island of Cunha, the Easter Islands, Brazil, China, Australia, Canada and the US.

THE IMMORTALIZATION OF CELLS

In order to have samples of the current human genome for the future, a bank of immortalized cell lines (lymphoblastoid cell lines = LCL) of individuals from various ethnic backgrounds from around the world was created at the Jean Dausset Foundation (CEPH) in Paris. The blood samples that served as sources for the LCL were freely donated under informed consent and collected by the CEPH and other laboratories working in the HDGP. For each sample —currently 1064 subjects from 51 communities from 60 geographic areas— the information is limited to sex, group and geographic area of origin. Only the donors of 47 populations are unrelated persons. The collection contains too many males, who are aploid for chromosomes X and Y. However, both men and women can be considered aploid for mitochondrial DNA, which is transmitted only from mother to offspring of both sexes. In this way, the phenotype pictures related to these genomes are equivalent to the genotypes. The LCL of samples with more males allow the collection of hundreds of SNP linked to chromosome Y, and thus to trace the history of the male half of the population; vice versa, the mitochondrial DNA markers specify the female history.

The DNA obtained from the LCL can be requested by researchers who agree to analyze all of them and return the results to the central database. The LCL cannot be released. So far, the variation in the DNA sequence has been studied at the level of 404 microsatellites related to the nuclear genome, including 20 from chromosome X and 7 from chromosome Y.

THE APE GENOME PROJECT

Towards the end of the 1990s, Japanese primatologists launched a research program to sequence the genome of apes which was called the Silver Project —from Ape Genome, whose initials AG are none other than the chemical symbol for silver. Soon thereafter, American scientists began a campaign to convince the public institutions to fund a similar study that looked only at chimpanzees, the ape that differs from humans by only 1.3% of its DNA., considering base substitutions (the difference raises up to 5% if we consider also the presence of indels). Half the difference accumulated after the two evolutionary lines diverged, and since our genome is composed of about 3 billion bases, about 20 million of them (0.65%) are mutated with respect to the African ape. Of these, less than 500,000 (2%) involve coding regions. So it is within this small block of unshared bases —coding and non-coding regions— that we have to look for what turned us into humans and them into chimpanzees. The comparative genome promises to become the most powerful instrument for acquiring data on functionally important conservative regions of DNA. It may also provide a key to the search for inter-specific differences, i.e. the basis for determining how the information stored in the double helix transforms into a phenotype.

The main stimulus to study the genome of apes comes from medicine, particularly from the need to understand why we humans are affected by certain diseases to which our nearest primate relatives seem to be immune. It is known, for example, that chimpanzees inoculated with HIV do not develop AIDS, and that although they share with humans a gene that predisposes to Alzheimer's disease, they do not manifest the disease. Determining the differences that characterize our genome could allow us to discover new therapies able to treat diseases claiming million of victims.

Fortunately, however, it is not only the interest in health at any cost that drives research forward. For many scientists, the Silver Project and the "chimpanzee project" provide the only way to shed light on the evolutionary relations between humans and other primates and on such issues as the complexity of the human mind. Of course, doubting skeptics have already voiced their opinions. One of them is Grahame Bulfield, the director of the Roslin Institute where the sheep Dolly was cloned. In his opinion, we would do well to remember that the small difference between us and chimpanzees means that we differ by many thousands of sequences and that to discover which of them have gone into determining the humanity of humankind and the "chimpanzeeness" of chimpanzees, like the susceptibility or immunity to certain diseases, will not be an easy task at all.

KEYWORDS Human. Genome Project. Human Diversity. Ape Genome

ABSTRACT The publication at the beginning of 2001 of the human genome project has opened up a new era into biological researches. Starting from these preliminary sequences, a complete list of all human genes and the proteins which they codify will be

filled out. This will permit their utilisation in biomedical investigations. Although the function of the largest part of the genome is still unknown (about 98%), it has been recognised that all the proteins and RNAs, which are involved in cellular functions, are codified by the picked out sequences. This outcome is of great importance because it will help identify all the molecular components and the control mechanisms of the cellular functions. The results of the Human Genome Project (HGP) together with the few obtained by the Human Genome Diversity Project (HGDP), which for some reason failed to attract the same widespread support and funding, and those gathered by the Ape Genome Project, will allow a better understanding of human evolution, the causes of diseases, and the relationships between genetic and environment in shaping the human condition.

REFERENCES

Cann, H.M.: A Human Genome Diversity Cell Line Panel. *Science,* **296:** 261-261 (2002).

Castiglioni, A.: Storia della medicina. A. Mondadori, Milano (1936).

Kevles, D.J. and Hood, L.: (a cura di). *The Code of Codes: Scientific and Social Issues in the Human Genome Project.* Harvard University Press, Cambridge, Mass. (1992).

Lewontin, R.: It *Ain't Necessarily So: The Dream of the Human Genome and Other Illusions.* NYREV, Inc., New York (2000).

Magner, L.N.: *A History of the Life Sciences.* Marcel Dekker, Inc., New York-Base (1979).

Marshall, E.: Gene Prospecting in Remote Populations. *Science,* **278:** 565 (1997).

Pennisi, E.: NRC OKs Long-Delayed Survey of Human Genome Diversity. *Science,* **278:** 568 (1997).

Redi, C.A., Zuccotti, M. and Garagna S.: L'«altro» genoma. *Le Scienze,* **409:** 36-42 (2002).

Vogel, F. and Motulsky, A.G.: *Human Genetics.* Springer-Verlag, Berlin-Heidelberg-New York-Tokyo, 2nd Ed. (1986).

Authors' Addressess: **Gianfranco Biondi,** Dipartimento di Scienze ambientali, Universita' di L'Aquila, Via Vetoio, 67010 L'Aquila, ITALIA
E-mail: gbiondi@tiscalinet.it
Olga Rickards, Dipartimento di Biologia, Centro dipartimentale di "Antropologia molecolare per lo studio del DNA antico", Universita di Roma "Tor Vergata", Via della Ricerca Scientifica 1, 00133 Roma, ITALIA
E-mail: rickards@uniroma2.it