THIS IS AN ADVANCED DRAFT OF A PUBLISHED PAPER.

REFERENCES AND QUOTATIONS SHOULD ALWAYS BE MADE TO THE PUBLISHED VERSION, WHICH CAN BE FOUND AT:

García-Sancho M. (2015) "Genetic information in the age of DNA sequencing", ***Information & Culture: A Journal of History,*** **50(1): 110-142.**

URL: http://dx.doi.org/10.7560/IC50105

Genetic information in the age of DNA sequencing

The claim that genes encode and transmit information is a central conceptual tenet in biomedicine. Historians have placed the origins of this claim in the rise of molecular biology after World War II and philosophers still debate the utility of understanding genes as information for biomedical research. In this paper, I will investigate how 'genetic information,' as both a concept and a model for experimental practice, was affected by the emergence, in the mid-1970s, of technologies which enabled scientists to determine the sequence of chemical units of DNA, the molecule which constitutes genetic material. I will argue that DNA sequencing, rather than changing the meaning of genetic information, transformed the possibilities of what could be achieved with this concept, and directed it to large-scale enterprises such as the Human Genome Project. Thus, my argument suggests that scientific concepts should be regarded as entities which, beyond abstract thinking, enable researchers to do things – in this case embarking in a multi-million project aimed at sequencing the information encoded in the human genome.

-1.Introduction:

Any biology textbook published over the last sixty years states, as a fact, that genes encode and transmit information. This statement derives from the mid-1940s, when biologists started to equate the activity of genes with a code which transmits instructions and governs all processes and features of living organisms, from respiration to eye color. The elucidation of the double helix of Deoxyribonucleic Acid (DNA) in 1953 provided genes with a precise molecular entity, and biologists subsequently established that genes exert their governing function by synthetizing proteins. Since then, the synthesis process has been described in terms of 'transcription' and 'translation' of information from DNA to proteins. More recently, the determination of the detailed sequence of chemical units in human DNA – the Human Genome Project (HGP), completed in 2003 – was described as the act of reading the 'book of life'.[1]

Insert Figure 1 around here (genetic code and DNA sequence)

Historians and philosophers of biology have long investigated the causes and implications of this informational thinking. The historical literature describes how, during the decade following World War II, a number of military-oriented lines of research exploring the accurate transmission of electric signals spread among civil society, including the life sciences. The influence of these lines of research led life scientists to address the mechanisms by which genes synthetize proteins as a "coding problem", and to adopt concepts and models from a range of wartime engineering fields, namely cybernetics, communication theory, and automata theory. Terms such as 'information transfer' or 'feedback' became widespread in biological research, and the metaphors of a telegraph or an electric circuit attempted to capture the action of genes in living organisms. Despite the lack of practical results of these analogies, the metaphors of genes as information systems diversified in the 1960s, setting up a conceptual framework that still persists today.[2]

The philosophical literature focuses on the current utility of the concept of genetic information and the consequences of using this type of metaphor in biomedical research. Over the last fourteen years, as the HGP has reached an end, an intense debate has emerged between those who defend the benefits of informational thinking and those who consider it an inadequate model for gene action. The defenders of genetic information argue that, despite their shortcomings, informational metaphors have inspired a long-standing line of research into DNA structure and function with innumerable valuable outcomes. The completion of the human genome sequence and earlier the determination of the mechanisms of protein synthesis – an achievement which was and is still described as the deciphering of the genetic code – are good examples of this information-led research. The critics of informational thinking claim that there are fundamental differences between the way genes function, and the natural or artificial languages that attempt to represent them. According to these scholars, approaching genes as information sets a misleading framework for understanding what to achieve with their investigation, this being one of the reasons why the HGP, and more generally recent genomic medicine, have not fulfilled their promises.[3]

Neither historians nor philosophers, however, have specifically addressed the connections between informational thinking and biomedical research in the period from the 1970s to 1990s. This seems to be a pressing task, given that in this period a number of crucially influential technologies emerged and were disseminated among biologists, namely recombinant DNA and sequencing. Sequencing enabled researchers to determine the precise linear structure – the sequence – of nucleotides that constitutes the DNA molecule. Recombinant DNA methods made it possible to modify the order of nucleotides within the

molecule, thus altering genetic material. Scholars have stressed how, under these techniques, 'reading' and 'rewriting' DNA became possible. These perceived possibilities, according to them, were crucial for the widespread support to the HGP and the concern that, under this massive sequencing enterprise, an irreversible blurring of the boundaries between nature, culture, and the human condition was being established.[4] Yet little is known about the links between these new potentialities of genetic information and the older work on the coding problem. In this paper, I will propose a genealogy that seeks, at the same time, to extend the historiography of genetic information beyond the 1960s and clarify where the informational metaphors that philosophers debate come from.

In my previous research, I offered a first approximation to this genealogy and claimed that, between the 1950s and 1990s, the notion of genetic information shifted from the idea of a message to the idea of a written text.[5] Now I want to argue that this transition was not so much triggered by a change of meaning in the concept of genetic information, but by this concept becoming embedded in a technology – DNA sequencing – that produced as an outcome a string of interconnected units – a nucleotide sequence. This sequence outcome, additionally, could be stored and read as data in a computer. The embedment of 'genetic information' in DNA sequencing provided this concept with an operative dimension that it lacked before. This operative dimension led biomedical researchers to address the information stored in the genome of different organisms, including humans, with the expectation of gaining insight into fundamental scientific and medical problems.

My argument will rest on two key theoretical contributions from the existing literature. Firstly, historians have demonstrated that genetic information was not a mere speculative idea: between the 1950s and 1960s, the concept was embedded in the experimental practices and techniques of biologists, and led to the use of viruses as Rosetta Stones, of gene expression systems as bearers of information and feedback, and of protein sequences as evolutionary documents.[6] I will argue that, with the advent of DNA sequencing, a new techno-experimental embedment of 'genetic information' provided this concept with previously unrealizable prospects and potentialities. Secondly, recent scholarship has documented the existence of a range of "scriptural analogies" in biology that preceded the emergence of the terms 'information' and 'code'. Among them, 'sequence', as a linear combination of chemical units which shaped biological molecules as letters shape a text, had a very influential role in the late 19th and early 20th century.[7] I will argue that, with DNA sequencing, both 'genetic information' and 'sequence' were for the first time embedded in the same technology, thus enabling researchers to equate the potentialities of this technology

with those of a written text. In the HGP and other large-scale sequencing projects, scientists – and later society – believed that 'genetic information' was actually written in DNA sequences that were referred to as the *book of life*.

I will show these processes of embedment by analyzing the careers of three UK-based researchers: Francis Crick, Sydney Brenner, and Frederick Sanger. Crick, the co-elucidator of the double helical structure of DNA, had a crucial role in defining the concept of genetic information as it was used by biologists. During the mid-1950s, he started co-working with Brenner and they both sought cooperation with Sanger for solving the genetic code problem. Sanger had just developed the first techniques for determining the sequence of chemical units which constitute protein molecules. Crick and Brenner expected to be able to use them for elucidating how DNA directed the synthesis of proteins. This cooperation did not lead to any visible results, but twenty years later, when Sanger extended his work to DNA in 1975, biomedical researchers increasingly considered DNA sequencing to be the preferred means for analyzing how genes convey information.

-*2.Information and sequence: the central dogma and the challenges of the coding problem*

The careers of Crick, Brenner, and Sanger have amply been addressed by scholars. Part of this literature specifically focuses on Crick's approach to the term 'genetic information'[8] and the cooperation between Crick, Brenner and Sanger in the context of the coding problem.[9] However, given the lack of immediate results of this cooperation, the scholarship has not explored the wider implications of Crick, Brenner and Sanger's attempt to work together. In the next two sections, I will first review what 'genetic information' meant for Crick and Brenner, and then analyze the long-term consequences of their collaboration with Sanger. I will argue that Sanger was an inspirational source in Crick's portrayal of DNA as a 'nucleotide sequence' in his papers following the elucidation of the double helix in 1953. Yet, Sanger's work had little impact on the way scientists addressed the transfer of information from genes to proteins up to the mid-1970s, when he invented the first techniques which enabled to handle DNA at the bench level.

2.1.Crick, information and molecular biology

Before addressing biological problems, Crick completed an undergraduate degree and started a PhD program in physics at University College London. The outbreak of World War II interrupted his doctoral studies and led to his mobilization at the UK Admiralty Research Laboratories, where he worked on the design of mines. After the War, he resumed his academic career and, like many physicists at that time, wanted to make a transition to biology. He joined the Cavendish Laboratory in Cambridge, an institution which traditionally had been devoted to physics, but since the 1930s was increasingly focusing on biological problems. Crick's work and new PhD project focused on X-ray crystallography, a technique that required considerable expertise in mathematics and allowed the determination of the three-dimensional structure of molecules. It was with this technique – and through collaboration with biologist James Watson, and other chemists and crystallographers – that Crick managed to model the double helical structure of DNA in 1953.

Crick has always described Edwin Schrödinger's *What is Life?* as a decisive reading for his move to biology.[10] Schrödinger was one of the founders of quantum theory and in this book, published in 1944, attempted to apply quantum mechanics to living organisms. In the second chapter of the book, entitled "The hereditary mechanism", Schrödinger stated that the chromosome fibers of the cell – which were known to be the molecules of which genes are made – contained "in some kind of code-script the entire pattern of an individual's future development". He also claimed that the chromosome code-script was, at the same time, the "architect's plan and builder's craft", given that it possessed the means for executing the developmental plan it contained.[11]

Watson and Crick used the term 'code', along with 'information' and 'sequence', to describe the properties of DNA shortly after the elucidation of the double helix. In a 1953 paper written months after the original double helix report, they stated that it was "likely that the precise sequence" of nucleotides in DNA was the "code" which carried "the genetical information".[12] This was the vision of the gene which started to prevail in the 1950s, given that the double helix and subsequent experiments proved that DNA was the constituent substance of the chromosome fibers and exerted its genetic function by synthetizing proteins. The view of DNA as an information carrier was consistent with Schrödinger's definition of the chromosome fibers as bearers of a code-script. Biologists subsequently focused on the code which transmitted – rather than on the sequence which carried – the genetic information and, during the remaining of the decade, were increasingly interested in elucidating the mechanisms by which DNA synthetized proteins. The synthetized proteins, in their turn, were

the catalysts of the chemical reactions involved in respiration, muscular contraction, and other essential organismal functions, including the replication and transcription of DNA.

Lily Kay has forcefully argued that the growing visibility of cybernetics, communication theory, and automata theory was crucial for the development of the terms 'code' and 'genetic information'. These fields, which Evelyn Fox Keller has collectively named "cybersciences", were mainly concerned with the accurate transmission of electric signals and, during World War II, played a crucial role in military projects such as radar or anti-aircraft artillery. Their proponents also attempted to apply communication engineering models to biological problems, mainly the brain sciences. The popularization of their views after 1945 led an increasing number of biologists to equate the mechanism of gene action – and later of protein synthesis – with an electric circuit: the performance of the correct organismal function was dependent upon the transmission of the right signal by genes and the reception of enough information to synthetize the appropriate protein.[13]

The link between this cyberscientific discourse and Crick was another physicist converted to biology, George Gamow. Involved in military planning as a consultant of the US Navy during World War II, Gamow saw protein synthesis as a problem of cryptanalysis, a field of mathematics engaged in the deciphering of codes. Cryptanalysis had experienced a substantial development during the War, and its techniques were refined by communication theory, especially the tools this field provided to quantify the minimum amount of information necessary to solve a code.[14] After reading the double helix paper, Gamow thought that protein synthesis could be portrayed as a cryptanalytic problem: if DNA sequences were formed by four different chemical units called nucleotides and helped synthetize protein sequences composed of 20 chemical units named amino acids, a mathematical analysis of the sequences would find patterns in the nucleotide or amino acid arrangements. From this information, it would be possible to attempt to deduce one sequence from the other.

In 1954, Gamow founded the RNA Tie Club, an informal gathering of researchers who regularly proposed cryptanalytic strategies to solve the code. Crick was from the beginning a member and, up to 1961, he postulated a number of coding schemes, all concerned with achieving unambiguous "readings" of DNA. If DNA sequences presented patterns which enabled one to know with certainty which nucleotides synthetized amino acids, it would then be feasible to deduce the coding mechanism and to match those nucleotide patterns with the resulting amino acids.[15] Crick found 288 possible unambiguous patterns within all possible combinations of nucleotides in the DNA sequence. This

unmanageably large number, together with the lack of applicability of other models by the Tie Club members, led to a gradual abandonment of the cryptanalytic strategy towards the end of the 1950s.

However, despite the lack of concrete solutions, the cryptographic models of the code helped Crick to postulate his adaptor and sequence hypothesis, as well as the central dogma of molecular biology. Crick proposed the adaptor hypothesis in 1955, stating that there was a molecule – later characterized as transfer Ribonucleic Acid (tRNA) – which mediated in the transmission of information from DNA to proteins. The sequence hypothesis and the central dogma were jointly published by Crick in a highly influential paper in 1958. The sequence hypothesis claimed that in the process of protein synthesis, the precise sequence of nucleotides in DNA determined the amino acid sequence of proteins, while the central dogma stated that the flow of information from DNA to the adaptor molecule to proteins was unidirectional: "once information has passed into protein, *it cannot get back again*".[16] This framework shaped – and to a large extent continues to shape – the experimental strategies of molecular biology, a nascent discipline of which Crick and his Cambridge colleagues were self-declared and accepted founders.

Insert Figure 2 around here (Crick's scheme of coding problem and central dogma)

Crick's notion of information – the determination of a protein sequence by a DNA sequence in a unidirectional synthesis process – did not strictly square, as scholars have argued, with that proposed by the cybersciences. For cyberscientists, information was a mathematical measurement that enabled them to quantify the transmission of a signal from one point to another. Claude Shannon, the founder of communication theory, stressed the abstract nature of 'information' by explicitly separating this notion from meaning: what the amount of information in an engineering system enabled was to assess the quality of transmission of a signal regardless of what was being transmitted. By contrast, the semantic dimension of information was crucial in Crick's postulations, given that what DNA transmitted via the adaptor molecule was the instructions to make a protein sequence. The idea of a genetic code thus entailed meaning or transition from a precise arrangement of nucleotides in DNA to a corresponding arrangement of amino acids in proteins. Crick himself has retrospectively highlighted these differences by explicitly denying any connection with Shannon and stating that the model that inspired the central dogma was the Morse code.[17]

This denial of a direct influence does not yet rule out a genealogy between the cybersciences' concept of information and that used in the emerging field of molecular biology. The connection with Gamow and the importance of cryptanalysis for the Tie Club, together with Crick's previous work in naval research and his skillful mathematical modeling, suggest that the formal and quantitative definitions of information that were circulating at that time shaped the formulation of the 1958 coding scheme, but in a rather loose way. On one hand, information for Crick and the molecular biologists who embraced the central dogma was a set of instructions encoded in the DNA molecule and transmitted by an adaptor – soon named as transfer RNA – much in the same fashion electric signals circulated in a circuit. On the other hand, the effect of those instructions – the determination of the synthesis of a precise protein sequence – was far more meaningful and specific than any other 'information' cyberscientists had worked with. The characterization of genetic information as a set of instructions acquired a life of its own following the abandonment of the mathematical models of the code and continued changing hand-to-hand with the experimental strategies of biomedical researchers.[18]

2.2.The connection with the phenotype and the entrance of Sanger's sequences

The absence of practical results with the cryptographic approach led molecular biologists to seek alternative strategies to address the coding problem. One of the alternative solutions was proposed by Sydney Brenner, a newly recruited scientist at the Cavendish Laboratory, who had arrived at Cambridge in 1956 and shared an office with Crick. Unlike Crick, Brenner's background was in medicine rather than physics, and he had previously worked on the genetics and biochemistry of bacteriophages – a type of virus that infects bacterial cells by inducing them to express their DNA. Bacteriophages had been an important model organism for locating genes since the early decades of the 20th century and were being increasingly adopted by molecular biologists in order to characterize cellular processes in the simplest possible life forms.

The previous work on bacteriophage gene mapping[19] served Brenner as the point of departure for his approach. Given that the approximate position of genes in some regions of the chromosomes of bacteriophages had been determined, Brenner submitted those regions to radiation, in order to produce mutations. Mutations are alterations in DNA sequences which lead to the production of correspondingly abnormal protein sequences. Brenner had the hope that by isolating the defective proteins, matching them with the mutated genes, and

comparing them to normal synthesis processes, an experimental approach to the coding problem could be established. By the late 1950s, molecular biologists increasingly regarded microorganisms as the new "Rosetta stones" to decipher the code and Brenner had been inspired by similar attempts developed in other viruses.[20]

Brenner, in consultation with Crick, considered that cooperation with researchers working on protein sequencing would strengthen his approach to the coding problem. A main representative of protein sequencing research was Sanger, who was based in the neighboring Department of Biochemistry of the University of Cambridge, and of whose work Crick was aware since the late 1940s. In 1949, Sanger had published a technique that enabled the determination of part of the amino acid sequence of the protein insulin. By the time he published the complete sequence, in 1955, Crick and Brenner persuaded him to engage in a collaborative project, with the hope of applying sequencing techniques to the bacteriophage proteins.[21]

Sanger's interests, like those of many biochemists at the time, had been exclusively focused on protein structure. Genetics did not appeal to him, and until his contact with Crick and Brenner, Sanger was unaware of the mathematical strategies employed to decipher the code. Sanger's background had been rather shaped by the work on the chemical nature of proteins, particularly that of German biochemist Emil Fischer between the late 19th and early 20th century. Fischer had demonstrated that proteins were formed by long chains of discrete amino acid elements, which were united to each other by chemical bonds. Building on this, Sanger had used his sequencing work as evidence against the hypothesis that the arrangement of amino acids in protein chains followed periodical intervals.

By the time of Sanger's early sequencing work, other biochemists had favored this 'periodicity hypothesis' in the arrangement of the protein chains and thus postulated that the location of amino acids could be mathematically predicted. If in a given protein chain, amino acid A was located at fixed intervals of four and amino acid B at three intervals, an equation taking into account these constraints could deduce the sequence. On the contrary, Sanger's first results on insulin showed that, in proteins, each amino acid could be followed by any other amino acid. This suggested that, in order to determine the sequence, it was essential to chemically analyze proteins rather than engage in mathematical predictions.[22]

Historian and philosopher Werner Kogge has shown that this concept of sequence, as a linear and undetermined combination of building blocks constituting biological molecules, was an influential "scriptural-notational structure" which "grew out of the interaction between different traditions of knowledge from the 1870s on". Late 19th century

embryologists first and early 20th century biochemists later used it to characterize, respectively, the process by which cells differentiate in the development of species from embryo to adult, and the basic structure of proteins. These biological researchers often compared the combination of chemical units in sequences with that of letters in a written text or of musical notes in a melody. In the 1950s, molecular biologists adopted the term 'sequence' and created a "single convolute" by superimposing it with the notions of 'code' and 'information'.[23]

Crick adopted the term 'sequence' in 1953, when in an article written immediately after the elucidation of the double helix he stated that the "precise sequence" of nucleotides in DNA was "the code which carries the genetical information".[24] Given that at that time he was closely following the work on protein sequencing, it is likely that he borrowed the term from Sanger's research. Crick and Brenner's cooperation with Sanger sought to bring to molecular biology the chemical expertise they lacked to analyze biomolecular sequences. If Sanger's technique was applied to the proteins resulting from Brenner's bacteriophage experiments and the scope of sequencing was expanded to the adaptor molecule (RNA) and DNA, it would be potentially possible to solve the genetic code by matching nucleotide and amino acid sequences. This coincided with an increasing interest by Sanger in the process of protein synthesis as a means to give continuity to his finished work on insulin by addressing other biologically related molecules.

The cooperation between Sanger, Brenner, and Crick coincided with plans to create an independent institution to house the growing group of molecular biologists at the Cavendish Laboratory. The Cavendish biological team was funded by the Medical Research Council (MRC), the body of the British Government managing biomedical research. In the face of the increasing importance and international prestige of the Cavendish biologists, the MRC decided to create the Laboratory of Molecular Biology of Cambridge (LMB) in 1962. That same year, Crick was awarded the Nobel Prize for the determination of the double helix, together with other Cavendish crystallographers. Sanger, who had also won the Nobel Prize in 1958 for his work on insulin, was invited to move to the LMB and head a division devoted to the continuation of his research on proteins, and the expansion of that work to RNA and DNA sequencing. He accepted the offer and moved together with his team from the Department of Biochemistry shortly after the LMB opened.[25]

One year before Sanger's move, in 1961, an alternative approach to the coding problem was presented by US-based biochemists Marshall Nirenberg and Heinrich Matthaei. This approach consisted of simulating the protein synthesis process in a test tube by

introducing synthetic RNA – the molecule which mediates between DNA and proteins – and analyzing the amino acid sequences that were formed. The synthetic RNA was intentionally very short and gave rise to correspondingly short amino acid sequences, whose determination did not require sequencing techniques. This approach soon proved the most efficient to decipher the code and, by 1967, researchers had matched all possible protein sequences with the RNAs used for their synthesis. Once scientists had established these matches, DNA sequences were easy to deduce, given that the synthetic RNA sequences were known in advance and DNA is molecularly very similar.

The effectiveness of Nirenberg and Matthaei's approach – and the fact that it did not involve sequencing – led the cooperation between Crick, Brenner and, Sanger to gradually lose intensity. During the 1960s, the three of them reoriented their interests, Crick initiating his research on neurosciences, which would prompt his move to the Salk Institute in 1976. Sanger left the sequencing of proteins to other members of his team and embarked in RNA sequencing, but without specifically linking this project to the coding problem. RNA, like DNA, is composed of nucleotide sequences. The shorter length of RNA and its suitability for chemical analysis had led other biochemists, prior to Sanger, to work on determining its sequence. Sanger chose two types of RNA for his sequencing experiments: transfer RNA (tRNA) – which had been identified as Crick's 'adaptor' – and messenger RNA (mRNA), which in 1961 was described by Brenner, Crick, and other researchers as a molecule also mediating in protein synthesis.[26] However, unlike in protein sequencing, the RNA efforts Sanger conducted did not include the same degree of collaboration and engagement with broader problems of molecular biology.

Brenner, in his turn, initiated a large-scale project on *C. elegans*, a small nematode worm of one millimeter length with an extremely short life cycle. In his project proposal, written in 1963, he stated that "all the *classical* problems of molecular biology" had either been solved or would be solved "in the next decade", mainly referring to the decipherment of the genetic code. He advocated for molecular biologists to address new research areas, and considered development and behavior as suitable future horizons.[27] Brenner proposed *C. elegans* as an ideal organism to investigate the genetic basis of these processes, given its simplicity and the short time span leading from embryo to adult state.

The first experiments on *C. elegans* followed the approach Brenner had established for bacteriophages. During the remainder of the decade, he obtained a genetic map of the worm's chromosomes, produced mutations in certain genes, and attempted to match the altered genes with developmental or behavioral anomalies in the worm. However, a key

difference from previous work was that Brenner no longer limited his research to protein products: he was attempting to connect the *C. elegans* genotype – mutated or normal genes – with its phenotype – observable organismic features.

Brenner published the first results of this work in 1974, in a lengthy paper in which he reported over ten years of complex crossings between normal and mutated worms. As a result of these crossings, he identified five genes responsible for behavioral and developmental alterations in the worm, which he coined as uncoordinated, roller, dumpy, small, and long. Brenner presented the experiments as a means to elucidate the "programme" which mediated between the worm's genes and its development and behavior. He defined development and behavior as the "result of a complex set of computations" that were initiated by genes and unfolded according to a "logical structure". By elucidating this logical structure in both mutated and non-mutated worms, Brenner expected to determine the mechanisms mediating between the transmission of genetic information and normal or abnormal phenotypic patterns.[28]

Brenner's hypothesis of a genetic programme built on the notion of information postulated by Crick, but he framed it in a different experimental and disciplinary culture. On one hand, the way the genetic programme operated squared with the definition of information as a unidirectional transfer of instructions that Crick had established in his 1958 central dogma paper. The overall objective of Brenner's project was, precisely, to deduce those instructions from their behavioral and developmental effects. On the other hand, Brenner's background in medicine rather than physics led him to investigate the flow of information via genetic experiments rather than mathematical modelling, first using bacteriophages and then the worm *C. elegans.* This medical training enabled him to follow information one step further and not limit its flow to the protein products of genes: Brenner's programme suggested a pathway for genetic information to circulate from *C. elegans* genotype to its phenotypic effects. For this pathway to be operational, Brenner needed to replace Crick's image of a telegraph Morse code by the more sophisticated notion of a programme, which rather than mere transmission involved some sort of transformation of the instructions into results. This notion of a 'genetic programme' was inspired by the computer apparatus which, at that time, were gradually permeating biomedical laboratories.[29]

Crick and Brenner's notion of genetic information by the mid-1970s was, thus, the result of a confluence of their experimental strategies and disciplinary backgrounds. Building on his undergraduate and wartime experience as a physicist, Crick had postulated a definition of genetic information as a set of instructions that were transmitted one-way from genes to

proteins and consequently could be mathematically traced. Brenner had used his interest in genetics to add to this definition a mechanism by which the genetic instructions could lead to phenotypic effects, much as a computer program produces an outcome by logically transforming a given information input. Their overall concept of information squared with Crick and Brenner's gene-centric view of biology and their research goal of determining the molecular mechanisms leading from DNA to proteins, and from proteins to behavioral and developmental effects.

Both Crick and Brenner were aware that genetic information was coded in a DNA sequence, this being the reason of their interest in Sanger's work. However, the absence of practical results of their cooperation led Crick and Brenner to blackbox the sequence, and focus their research on determining how genetic information was transmitted. This situation changed substantially with the advent of Sanger's DNA sequencing techniques in 1975, one year after the publication of the first results of Brenner's *C. elegans* experiments. The availability of a technology which could determine DNA sequences transformed the way researchers approached the phenomenon of genes encoding and transmitting information, as well as their expectations of what to achieve with this information.

-3.Information as sequence: the impact of Sanger's DNA techniques

When Brenner, Crick, and Sanger started their cooperation in the mid-1950s, DNA sequences were the object of much discussion but little experimental practice. Crick and Brenner considered these sequences as an essential component in the transmission of genetic information, and expected that Sanger's work would enable them to determine how this information was expressed and circulated from DNA to RNA to protein sequences. Yet, the gradual decrease in their cooperation led DNA sequences to remain as abstract entities, invoked in papers but never tackled through the application of technologies. In 1975, with the arrival of Sanger's first DNA sequencing techniques, DNA sequences were, for the first time, embedded in a technology that produced tangible experimental results.

In the next two sections, I will argue that DNA sequencing did not substantially alter the way researchers understood 'genetic information'; DNA sequencing rather affected how researchers translated this concept into experimental practice and what they expected to achieve with the experimental results. The biomedical community at large continued to accept Crick's definition of 'genetic information' as a set of instructions that were transmitted

one-way from genes to proteins and from proteins to the phenotype. However, with DNA sequencing, this information became a goal in itself rather than a means for determining the genetic code or the genetic programme. As DNA sequencing techniques spread in the 1980s, molecular biologists increasingly opted to determine genetic information in the form of DNA sequences rather than deducing it from its protein products or phenotypic effects; they did so with the expectation that the genetic information encoded in the DNA sequences would, conversely, enable them to deduce protein products and phenotypic effects.

3.1. DNA sequencing and Sanger's engagement with genetic information

Towards the end of the 1960s, having sequenced a number of RNA molecules, Sanger decided to start tackling DNA. The length of this molecule, with considerably more chemical units than proteins and RNA, led him to modify his sequencing strategy and design a method that combined chemical and biological approaches. Sanger's identity as a biochemist had dominated his protein and RNA sequencing techniques, which were framed in the experimental culture of analytical chemistry. In these techniques, the molecule was broken into fragments that were then submitted to the action of different reagents, in order to identify the amino acid or nucleotide units. In DNA sequencing, Sanger copied rather than degrade the molecule, and deduced the sequence from the copying process (see figure below). This approach of letting biological processes run and observing their effects rather than intervening in their course had distinguished the experimental strategies of the LMB as opposed to biochemical laboratories, such as Sanger's previous home institution. However, Sanger's DNA techniques maintained some of the original biochemical identity of sequencing in the use of *in vitro* instead of *in vivo* methods, and the necessity of structural chemistry, in order to assemble the gradually determined nucleotides into a whole sequence.[30]

Another intrinsically biochemical feature of DNA sequencing was Sanger's research goal and the way he used the technique to achieve it. The concept of sequence, as Kogge has shown, dates back to the late 19th and early 20th century cultures of embryology and protein chemistry. As Sanger used it, this sequence concept involved determining the fine chemical structure of DNA with a combined approach of biology and analytical chemistry. This use of the term 'sequence' differed from the term molecular biologists had developed during the preceding decades. When in 1953, Watson and Crick stated that the "precise sequence" of nucleotides in DNA was the "code" which carried "the genetical information" – and Crick and other molecular biologists repeatedly defined DNA as a "sequence" in their subsequent

papers on the genetic code – their emphasis was on the code rather than the sequence. They acknowledged the sequence only as a carrier of the genetic information that shaped the synthesis of proteins – and later, with Brenner's postulation of a genetic programme, the deployment of normal or abnormal phenotypic patterns.[31] Sequences, for Sanger, were ends in themselves and, under his techniques, the connection with codes or programmes could only be established once the nucleotides had already been determined and assembled.

Sanger's first DNA technique was called the plus and minus method, and was published in 1975. Shortly after this, he presented the method to the Royal Society of London through the prestigious Croonian Lecture. Sanger's presentation started with a general description of DNA as genetic material and its role in living organisms. His description built on the language of information, but incorporating a subtle difference with regard to its previous use by molecular biologists:

> DNA, the chemical component of the gene, plays a central role in biology and contains the whole information for the development of an organism coded in the form of a sequence of the four nucleotide residues. The lecture describes the development and application of some methods that can be employed to deduce sequences in these very large molecules.[32]

Sanger's 'information' was still a set of genetic instructions encoded in a DNA sequence and with the capacity of shaping the development of the organism. However, the emphasis of his lecture was not on the instructions the information coded for – as it had been the case in Crick and Brenner's papers – but on the DNA sequence which embodied that code. Furthermore, Sanger was presenting *methods* that enabled researchers to determine those sequences, building on the techniques of analytical chemistry and combining them with molecular biology, his new institutional and disciplinary home. Those methods contrasted with the previous experimental strategies of molecular biology, framed more in mathematical physics and genetic crossing. The way Sanger presented his work, when compared with research previously conducted at the LMB, raised an underlying conclusion that was similar to that of the preceding protein sequencing techniques: information needed to be chemically analyzed – by determining molecular sequences – instead of mathematically computed – as in Crick and Brenner's attempts to deduce genes from their protein products or phenotypic effects.

Sanger's team used the plus and minus technique to determine the sequence of PhiX-174, a bacteriophage virus of the same type of those Brenner and other molecular biologists had employed in their genetic code experiments. The team confirmed the viral sequence with

the dideoxy technique, a more efficient method Sanger devised in 1977. Due to the extension of the sequence of PhiX-174 – of various thousands of nucleotides – the storage of the data gathered in each sequencing experiment and the assemblage of that data in a single sequence required the adoption of computers.

Sanger's team first used mainframe apparatus, which were physically large and external to laboratories, being operated at centers of calculation via punched cards. The computers' functions were based on a simple input-output mechanism, in which researchers submitted to the centers of calculation the cards with the information, and the mainframe operators, after processing them, returned the output results. Cambridge molecular biologists had long used these apparatus to perform the calculations needed to determine the three-dimensional structure of proteins, and Brenner had modeled on them his notion of a genetic programme mediating between *C. elegans* genes, and the worm's development and behavior. However, the researchers at Sanger's laboratory soon shifted to minicomputers, smaller devices that biomedical institutions had started introducing during the 1960s.[33]

Minicomputers were shared pieces of equipment located in common rooms. Users could directly operate the minicomputers, which did not require the submission of punched cards to external centers of calculation. This had led to an increased use of these devices as text processors and the emergence, during the 1970s, of word processing software. The text-processing feature of minicomputers served well both as a tool for the DNA sequencing projects and as a technology on which the nucleotide sequences, as bearers of genetic information, could be modeled. Genetic information was, thus, no longer a Morse code operated by a telegraph or a program to be run in a mainframe computer: it was rather a string of nucleotides that could be compiled, processed, and stored in an interactive minicomputer.[34]

The new association of genetic information with a molecular sequence that could be both determined and stored mirrors Lenny Moss's concept of Gene-D. In a historical and philosophical essay on how the contemporary concept of the gene has emerged, Moss identifies Genes-D as nucleotide sequences that set a developmental framework for the organism to unfold all its inborn genetic potentialities. The realization of these potentialities depends on a number of factors that are external to genes and interact with them in the process of development from embryo to adult – for instance, life habits, weather conditions, or other environmental contingencies. Moss counterpoises this concept to Gene-P, a preformationist as opposed to developmental gene, which represents a given phenotypic feature. Genes-P are not defined by their nucleotide sequence, but by their phenotypic effects, being referred to as the genes for blue eyes, albinism, or other organismic features. While

Genes-D are indeterminate regarding their consequences in the organism – these consequences depend upon other contingent factors – in Genes-P the contingent factors are blackboxed and researchers, for the sake of practicality, assume a straight connection between genotype and phenotype. In order to counter-balance their assumption, these researchers should never equate Genes-P with a nucleotide sequence, given that the transition between the sequence – Gene-D – and a phenotypic effect is mediated by environmental elements outside the genes.

The gene, as both a concept and a research object, can behave like both a Gene-P and a Gene-D. Researchers, according to their interests, may approach genes in either way, but need to be clear that genes do not function, at the same time, as Genes-D and P; in other words, DNA sequences do not directly determine phenotypic traits. Moss identifies Gene-P with a strong preformationist tradition initiated in the 17th century and which studied the transmission of hereditary characteristics from one generation to the other (preformationism means the belief that the embryo corresponds with a miniature representation of the features of parents). Gene-D, by contrast, was framed in embryological research and addressed the gradual transformation that organisms undergo from embryo to adult. Both traditions were interconnected up to the emergence of genetics as a discipline in the early 20th century. The first, classical geneticists, decided to focus exclusively on the hereditary transmission of features – Gene-P – and leave aside how these inherited features unfolded over life course – Gene-D. Since then, according to Moss, the gene represented by Gene-P has gained momentum and informed the emergence of molecular biology, while Gene-D, up to very recently, was relegated to the periphery of the life sciences – in the form of research into epigenetics or developmental biology.[35]

Crick and Brenner's investigations – and more generally the early work of molecular biologists – were informed by a Gene-P view. Their interests were in the effects of genes over protein synthesis – Crick's research on the genetic code – and later over phenotypic features – Brenner's project on *C. elegans.* Consequently, the central dogma and later the notion of a genetic programme emphasized the connections between genes, proteins, and the phenotype in the form of a one-way transmission of information. Both Crick and Brenner stated that a DNA sequence produced this information, but their conceptual and experimental frameworks blackboxed the sequence in favor of a focus on the logical structure of information transmission, with the hope of deducing the input genes from their protein products or phenotypic effects. Despite Brenner addressing the problem of development in his *C. elegans* research and presenting this concern as a re-embracement of development by

genetics research, he regarded the worm's transition from embryo to adult as an exclusive effect of gene action. In a 1973 paper, he explicitly acknowledged that the problem of how behavior was mediated by extra-genetic factors was "an entirely separate question at the moment".[36]

The concern of Sanger with DNA sequences and the connection he made between those sequences and development in his Croonian Lecture may be seen as more in line with the Gene-D concept. However, the rest of the lecture suggests instead that Sanger plainly assimilated the notions of 'gene' and 'information' as formulated in the central dogma. Firstly, in the abstract of the lecture Sanger stated that DNA contains "the whole information" for the development of an organism, implying that he regarded development as a phenomenon completely mediated by genes, in line with Brenner's simultaneous research on *C. elegans*. Secondly, Sanger started his sequencing experiments from a genetic map of PhiX-174, just as Brenner had done at the beginning of the worm project. And thirdly, towards the end of the lecture Sanger attempted to link certain regions of the DNA of PhiX-174 – those corresponding to genes – with protein sequences, favoring the assumption of a straight connection between both molecules.[37] There is, still, certain ambiguity in Sanger's view of genes and the information they transmitted, given that he never explicitly stated how those entities squared with his newly devised DNA sequencing techniques.

3.2.Reading, reading off and the contested limits of sequences

Both Sanger's plus and minus, and dideoxy methods, produced as an outcome a number of DNA fragments, which were derived from copying the molecule to be sequenced. The fragments were marked with a radioactive substance and separated on a gel, so they appeared as dark bands when the gel was photographed after separation. This photograph was called an autoradiograph and reflected a pattern of black bands corresponding to the DNA fragments. A trained researcher was able to determine the sequence by analyzing the position of the bands within the gel. In all his DNA papers, Sanger referred to this practice as "reading off" the sequence from the autoradiograph.[38]

Insert Figure 3 around here (scheme of copying technique and autoradiograph)

Unlike Crick and Brenner, Sanger was a scientist exclusively concerned with bench work and not prone to theoretical speculations. His papers just describe the experiments and he always

refrained from discussing too broadly their significance. This lack of background discussion makes it difficult to know what Sanger exactly implied with the expression "reading off": he never explicitly defined the term in his papers. The Oxford English Dictionary (OED) provides three definitions for 'read off' 1) the action of taking "a reading of (a measurement, value, etc.) from an instrument"; 2) deriving "(a result, conclusion, etc.) directly from tabulated data", and 3) reciting or recording "(items in a list, etc.) in sequence". These definitions contrast with a biological entry for the term 'read' in the OED, meaning "to interpret or extract genetic information from (a particular nucleic acid sequence)", especially "during the process of genetic transcription or translation".[39]

Sanger's use of the term 'read off' seems consistent with the OED definition, as the practice of extracting data – a DNA sequence – from a scientific tool – the autoradiograph – and recording the data in a notebook or computer program as a list of As, Cs, Ts and Gs – the abbreviations for adenine, cytosine, thymine and guanine, the four DNA nucleotides. By contrast, the OED quotes Crick's 1957 and 61 papers on the genetic code as examples of the biological definition of the term 'read': in those papers, reading the DNA sequence meant extracting nucleotide arrangements which coded for an amino acid in the process of protein synthesis. That action could either be performed by the genetic machinery or by the researcher attempting to solve the code mathematically and looking for patterns which could code for amino acids within all possible nucleotide combinations in the DNA sequence (see above).

This distinction between 'reading' and 'reading off' was not so marked among other biomedical researchers. Walter Gilbert and Allan Maxam belonged to a younger generation of molecular biologists at Harvard University, and devised an alternative DNA sequencing method in 1977, which rivaled Sanger's up to the mid-1980s. Maxam and Gilbert's method also produced an autoradiograph as an outcome. However, in the papers presenting the technique, the authors referred to the practice of extracting the sequence from the band pattern of labeled DNA fragments as just "read", never using the term 'off'.[40]

In 1980, Sanger and Gilbert were awarded the Nobel Prize for their DNA sequencing methods. During the Nobel Lectures, Sanger continued describing the act of extracting the sequence as 'reading off' the autoradiograph, while Gilbert kept using the term 'read' and, at some points in the discussion, speculated with the potential that DNA sequences could be granted accordingly:

> DNA is the information store that ultimately dictates the structure of every gene product, delineates every part of the organism. The order of the bases along DNA contains the complete set of instructions that make up the genetic inheritance. We do not know how to interpret those instructions; like a child, we can spell out the alphabet without understanding more than a few words on a page.

Later in the lecture, when Gilbert expressed his vision of the future of DNA sequencing, he became less cautious about human and biological reading capacities:

> We cannot read the gene product directly from the chromosome by DNA sequencing alone. Nonetheless (…) the hope exists, that as we look down on the sequence of DNA in the chromosome, we will not learn simply the primary structure of the gene products, but we will learn aspects of the functional structure of the proteins – put together over evolutionary time.[41]

Gilbert's remarks suggest that the perceived potential of DNA sequences was exponentially increasing as his technique – and Sanger's – spread among biomedical researchers. Sequence information was being recast from data that was read off autoradiographs to a royal road towards protein function and, therefore, a privileged means to deduce phenotypic attributes in the organism. This shift was concomitant with the growing analogy researchers established between the practice of reading a DNA sequence and that of reading a text: when the reader is familiar with the language of a written text, the meaning can be inferred automatically as the words are scanned with the eye. Gilbert forecasted that this would happen with autoradiograph band patterns as soon as researchers learned the language of DNA. With the transition from 'reading off' to 'reading', meaning or interpretative power was being added to the act of determining DNA sequences: researchers would potentially understand those sequences as the genetic machinery interpreted them in the process of protein synthesis.

This written-text analogy acquired additional ramifications as biomedical researchers introduced the new techniques into the emerging field of genomics. The journal *Genomics* was first printed in 1987 as the reference publication for "the newly established discipline of mapping / sequencing" DNA. Its first editorial stated that "the ultimate map, the sequence, is seen as a rosetta stone from which the complexities of gene expression in development can be translated and the genetic mechanisms of disease interpreted".[42] This use of 'rosetta stone' contrasted with that of the early experimental approaches to the genetic code in the late 1950s. While Brenner and other molecular biologists had placed the interpretative power of translating genetic information in the protein synthesis mechanisms of simple organisms –

such as bacteriophage viruses – the new genomic scientists were moving this potential to DNA sequences alone.

With the launch of the HGP in 1990, the power of genetic information was pushed to the realms of medicine and biomedical practice. In one of the first popular books on that Project, Gilbert predicted that a "theoretical biology" would emerge as a "science of pattern recognition", consisting in the extraction "from the genome sequence" of "the identity of human genes, their interrelationships, and their control elements", in order to deduce "how the genes and their proteins function". Leroy Hood, another main advocate of the HGP, claimed that the resulting human DNA sequence would lead medicine to move "from a reactive mode (curing patients already sick) to a preventive mode (keeping people well)" by virtue of analyzing the sequences of genes involved in hereditary diseases.[43]

This new notion of genetic information as, at the same time, a DNA sequence and a royal road to the understanding of protein function squares with Moss's argument of a conflation between Gene-D and P. For Moss, as the 20th century advanced, researchers increasingly superimposed and confused the meaning of these two independent gene concepts, thus considering genes as, simultaneously, DNA sequences – Genes D – and determinants of phenotypic features – Genes P. This has led to an over-simplistic view of the connections between genes, diseases, and personal attributes, with researchers overlooking the extra-genetic factors that mediate in the transition between a given DNA sequence and a phenotypic trait. The above discussed episodes, with Brenner considering the non-genetic determination of *C. elegans* behavior as a separate question or *Genomics* describing DNA sequences as rosetta stones, seem to square with this conflated view of the gene.

Moss traces the origins of this conflation to the separation, in the early 20th century, of the study of development from the investigation of hereditary transmission, the latter being chosen as the object of the new discipline of genetics. However, a turning point was the emergence of the informational jargon characteristic of molecular biology after World War II. The reading of Schrödinger's 1944 book – which attributed to the hereditary code-script the role of architect's plan and builder's craft at once – provided molecular biologists with a framework to present their research as the way of explaining, "in physicochemical terms, how the genotype contains within itself the instructions for making an organism". The HGP and the emergence of genomics represented, for Moss, the "culmination" and, at the same time, the "exhaustion" of this research strategy.[44]

My story suggests that the effects of this informational jargon and the conflated view of the gene it enhanced were limited to the conceptual level up the late 1970s. Despite Crick,

Brenner, and possibly later Sanger believing in a straight connection between DNA sequences and phenotypic traits, determining those sequences and deducing those traits remained as independent goals until the plus and minus, and dideoxy methods, were invented. It was these methods that enabled both Gene-D and P, DNA sequences and genetic information, to be embedded, for the first time, in the same technology. This technological embedment led the conflated view of genetic information to be mobilized not just via a gene-centric discourse,[45] but also via specific laboratory practices and DNA sequencing projects.

The spread of sequencing techniques, and particularly how their 'DNA reading' potential was both conveyed to and received by biomedical researchers, resulted in the HGP and other large-scale genomic initiatives to be perceived as projects addressing, at the same time, genetic sequences, their protein products and the phenotypic effects they produce in the organism. The presentation of the human genome sequence as "the book of life" in 2000[46] was, thus, the result of a complex informational genealogy in which concepts were inextricably linked to disciplinary trajectories, experimental practices, and technological regimes. Due to this inextricable linkage, the analysis of genetic information in the age of DNA sequencing entails more than an abstract discussion of ideas.

-6.Conclusions:

The establishment of a genealogy that starts with the first references to genes as information by biomedical researchers and finishes with the determination of the human genome sequence (1940s to 2000s) provides two fundamental insights to the way the concept of genetic information is approached in the academic literature. Firstly, the genealogy highlights that there is a considerable gap between the historical work on the origins of informational thinking in biomedicine – with the authors largely stopping their research in the 1960s – and the philosophical debates on the utility of the term 'information' in the age of genomics. Secondly, the genealogy demonstrates that the main transformation the genomic age introduced was not so much in the meaning of 'genetic information', but in the possibilities of what could be done and achieved with this concept.

Biomedical researchers long believed in a straight linkage between genetic sequences and phenotypic traits, but before the advent of DNA sequencing there was no possible way to connect sequences and the study of the transmission of genetic information at the laboratory bench. Sanger's techniques and the way they were received by molecular biologists – as reading devices – enabled to embed in the same technology two long-standing scriptural

analogies in biomedicine: molecular sequences and genetic information. This provided both analogies with a new operational dimension and while in the 1950s-to-1970s information had been a means to investigate gene action, throughout the 1980s it became a goal embodied in projects which produced DNA sequences as outcomes.

This historical transformation shows that the meaning and potentialities of genes as information are not only – nor mainly – the result of abstract thinking: throughout the second half of the 20th century, 'information' has been rather a by-product of the strategies researchers deployed for handling genetic material at the laboratory bench. And these strategies were, in turn, shaped by the disciplinary background of researchers and the technologies surrounding them. In other words, the history of the concept of genetic information can be written as a transition from the practices of physics and genetics to those of analytical chemistry, and from the telegraph and mainframe apparatus to the minicomputer as the technology to aid those practices. This connection between concepts, practices and technologies – and not just metaphysical ideas – should be taken into account in the growing debates about the utility and possible alternatives to the current informational definition of the gene.[47]

When all the relevant dimensions are considered, one can see that the conflation of meanings that genetic information entails today is the consequence of an underlying conflict of scientific goals. Genes are perceived, at the same time, as DNA sequences and determinants of phenotypic traits because scientists believed – and to a large extent still believe – that the old objective of genetics – deducing how genes work – could be directly achieved by the new sequencing techniques, connected to the power of computers. The slow fulfillment of this expectation after the conclusion of the HGP has led biomedical researchers to establish more sophisticated ways of linking genotype and phenotype. Nonetheless, the hope still exists that finding a connection between sequences and gene function is just a matter of time, effort and money.

This paper suggests that historical vision, as much as scientific and socio-political debate, is necessary to avoid disappointment. The separation between sequencing and the investigation of gene function should be re-instated, in order to show that DNA sequences may help deduce phenotypic effects, but are not the royal road to them. Research on biomolecular sequences and on phenotypic traits derives from different scientific traditions, and only a historical contingency led to their grouping under the umbrella term 'genetic information'. It is now time, forty years after this contingency, to move on and see how

current disciplinary, experimental, and technological interactions provide genetic information with a necessary new meaning.

Acknowledgements

Special thanks are given to Hans-Jörg Rheinberger, David Bloor, Fernando Broncano and Lenny Moss for insightful comments on a previous draft of this article. The journal editors kindly accelerated and eased the copy-editing process, so that the manuscript was published in a record time. This paper is the result of a long-standing line of research which began in 2004, when I was a PhD student at the Centre for the History of Science, Imperial College, London. It has benefitted from interactions with numerous colleagues and institutional support at the Residencia de Estudiantes, University of Manchester, Spanish National Research Council (CSIC) and currently the University of Edinburgh.

A detailed list of all sources of support would be unfeasible. However, I would like to highlight funding from a Hans Rausing Fellowship, Chancellor's Fellowship and various Spanish State-funded programs, among them the JAE-Doc, Juan de la Cierva and the projects HUM2006-04939/FISO, FFI2009-07522 and FFI2012-34076.

[1] See Francis Collins's speech – then director of the HGP – during the presentation of the first draft of the human genome sequence, in 2000. The ceremony was chaired by US President, Bill Clinton, and attended by the CEO of Celera Genomics, Craig Venter, and the UK Prime Minister, Tony Blair, who participated via satellite. http://clinton5.nara.gov/WH/New/html/genome-20000626.html (last accessed June 2014).
[2] Lily Kay, *Who Wrote the Book of Life: A History of the Genetic Code* (Stanford: Stanford University Press, 2000). Jerome Segal, *Le Zéro Et Le Un. Histoire De La Notion D'information Au Xxe Siècle* (Paris: Editions Materiologiques, 2011), Ch. 7. Evelyn Fox Keller, *Refiguring Life: Metaphors of Twentieth Century Biology* (New York: Columbia University Press, 1995). Sahotra Sarkar, "Biological Information: A Sceptical Look at Some Central Dogmas in Molecular Biology," in *The Philosophy and History of Molecular Biology: New Perspectives*, ed. Sahotra Sarkar (Dordrecht, Netherlands: Kluwer, 1996), 187-233.
[3] John Maynard Smith, "The Concept of Information in Biology," *Philosophy of Science* 67, no. 2 (2000), 177-94. Paul E. Griffiths, "Genetic Information: A Metaphor in Search of a Theory," *Philosophy of Science* 68, no. 3 (2001), 394-412. Ulrich Stegmann, "Genetic Information as Instructional Content," *Philosophy of Science* 72, no. 3 (2005), 425-43. A. Tauber and S. Sarkar, "The Human Genome Project: Has Blind Reductionism Gone So Far?," *Perspectives in Biology and Medicine* 35, no. 2 (1992), 220-35; Lenny Moss, *What Genes Can't Do* (Cambridge: MIT Press, 2004).
[4] Hans-Jorg Rheinberger, "Beyond Nature and Culture: A Note on Medicine in the Age of Molecular Biology," *Science in Context* 8, no. 1 (1995), 249-63. Barry Barnes and John Dupre, *Genomes and What to Make of Them* (Chicago: University of Chicago Press, 2008). Timothy Lenoir, "Makeover: Writing the Body into the Posthuman Technoscape," *Configurations* 10, no. 2 (2002), 203-20; Adam Bostanci, "A Metaphor Made in Public," *Science Communication* 32, no. 4 (2010), 467-88.
[5] Miguel García-Sancho, "The Rise and Fall of the Idea of Genetic Information (1948-2006)," *Genomics, Society and Policy* 2, no. 3 (2007), 16-36.
[6] Christina Brandt, "Genetic Code, Text, and Scripture: Metaphors and Narration in German Molecular Biology," *Science in Context* 18, no. 4 (2005), 629-48; Hans-Jorg Rheinberger, "The Notions of Regulation,

Information, and Language in the Writings of François Jacob," *Biological Theory* 1, no. 3 (2006), 261-67; Edna Suárez-Díaz, "The Rhetoric of Informational Molecules: Authority and Promises in the Early Study of Molecular Evolution," *Science in Context* 20, no. 4 (2007), 649-77.

[7] Werner Kogge, "Script, Code, Information: How to Differentiate Analogies in the 'Prehistory' of Molecular Biology," *History and Philosophy of the Life Sciences* 34, no. 4 (2012), 603-35.

[8] Lily Kay, 2000, ch.3. Evelyn Fox Keller, 1995, 94 and ff. Lindley Darden, "Flow of Information in Molecular Biological Mechanisms," *Biological Theory* 1, no. 3 (2006), 280-87. Paul E. Griffiths and Karola Stotz, *Genetics and Philosophy: An Introduction* (Cambridge: Cambridge University Press, 2013), ch.3. Gregory Morgan, "Early Theories of Virus Structure," in *Conformational Proteomics of Macromolecular Architectures*, ed. R Holland Cheng and Lena Hammar (Singapore: World Scientific, 2004), 1-40. Soraya de Chadarevian, "Of Worms and Programmes: Caenorhabditis Elegans and the Study of Development," *Studies in History and Philosophy of Biological and Biomedical Sciences* 29, no. 1 (1998), 81-105.

[9] Soraya de Chadarevian, "Sequences, Conformation, Information: Biochemists and Molecular Biologists in the 1950s," *Journal of the History of Biology* 29, no. 3 (1996), 361-86; Miguel García-Sancho, "A New Insight into Sanger's Development of Sequencing: From Proteins to DNA, 1943–1977," *Journal of the History of Biology* 43, no. 2 (2010), 265-323: 288 and ff.

[10] Robert Olby, *Francis Crick: Hunter of Life's Secrets* (New York: Cold Spring Harbor University Press, 2009), ch.4; R. Olby, *The Path to the Double Helix: The Discovery of DNA* (Dover Publications, 1994), section V. Francis Crick, *What Mad Pursuit: A Personal View of Scientific Discovery* (London: Weidenfeld and Nicolson, 1988), 18 and ff.

[11] Erwin Schrodinger, *What Is Life?* (Cambridge: Cambridge University Press, 2002 [1944]), 21-22. Historians have questioned the relevance of *What is Life?* in Crick and other physicists' transition to biology. Even though the influence of this book may have been exaggerated retrospectively, it seems plausible that Schrödinger's notion of a 'code script' had an important role, specifically, in Crick's engagement with the coding problem and its conceptualization as a transfer of information from genes to proteins. On the questioning of Schrödinger's influence see Pnina Abir-Am, "Themes, Genres and Orders of Legitimation in the Consolidation of New Scientific Disciplines: Deconstructing the Historiography of Molecular Biology," *History of Science* 23, no. 1 (1985), 73-117: 101 and ff.

[12] James Watson and Francis Crick, "Genetical Implications of the Structure of Deoxyribonucleic Acid," *Nature* 171(1953), 964-67: 965.

[13] Lily Kay, 2000, ch.3. See also Evelyn Fox Keller, 1995, 85 and ff.

[14] Lily Kay, 2000, ch.4. Communication theory enabled to mathematically quantify the amount of information of an encrypted message. With this measure, it was possible to estimate the redundancy – unnecessary information to accurately receive the message – and use it to either decipher an enemy code or make your own code non-redundant and thus impossible to break.

[15] Francis Crick, Leslie Barnett, Sydney Brenner, and Richard J. Watts-Tobin, "General Nature of the Genetic Code for Proteins," *Nature* 192(1961), 1227-32; Francis Crick, John S. Griffith, and Leslie Orgel, "Codes without Commas," *Proceedings of the National Academy of Sciences of the US* 43, no. 5 (1957), 416-21. Crick's idea of unambiguity within the DNA sequence was based on the assumption – later found correct – that each protein amino acid was synthetized by three nucleotides – these nucleotide arrangements being called triplets or codons. Therefore, if there were four different nucleotides in the DNA sequence and the nucleotide triplets ABC and CDA synthetized amino acids, then BCC and CCD would be unviable combinations, because they would lead to misleading "readings". See also Lily Kay, 2000, 163 and ff.

[16] Francis Crick, "On Protein Synthesis," *Symposia of the Society for Experimental Biology* 12(1958), 138-63: 153. The adaptor hypothesis was circulated in an informal note written by Crick to the other members of the Tie Club in 1955, see Lily Kay, 2000, 165 and ff. On the sequence hypothesis and the central dogma see Lindley Darden, 2006: 281-2. See also a special issue of *History and Philosophy of the Life Sciences* 28, no. 4 (2006).

[17] Lindley Darden, 2006: 285 and ff. Evelyn Fox Keller, 1995, 89 and ff. Crick's denial of any influence by Shannon and preference for the Morse code was stated in a letter he sent in 1998 to philosopher of biology Gregory Morgan. I wish to thank his generosity in sharing the letter with me.

[18] Lily Kay, 2000, 178 and ff.

[19] On the early genetic mapping work on bacteriophages see Frederic Lawrence Holmes, *Reconceiving the Gene: Seymour Benzer's Adventures in Phage Genetics* (New Haven: Yale University Press, 2006). Neeraja Sankaran, "Mutant Bacteriophages, Frank Macfarlane Burnet, and the Changing Nature of "Genespeak" in the 1930s," *Journal of the History of Biology* 43, no. 3 (2010), 571-99. William Summers, "Bacteriophage Research: Early History," in *Bacteriophages, Biology and Applications*, ed. Elizabeth Kutter and Alexander Sulakvelidz (Boca Raton, Florida: CRC Press, 2005), 5-27.On Brenner's early career, see Sydney Brenner, *My Life in Science* (London: BioMed Central, 2001), chs. 1-3.

[20] Lily Kay, 2000, 177 and ff. For similar experiments on the Tobacco Mosaic Virus see Christina Brandt, 2005: 638 and ff. Angela Creager, *The Life of a Virus: Tobacco Mosaic Virus as an Experimental Model, 1930-1965* (Chicago: University of Chicago Press, 2002), 298 and ff.
[21] Robert Olby, *Francis Crick: Hunter of Life's Secrets,* 2009, 125-6; Francis Crick, *What Mad Pursuit: A Personal View of Scientific Discovery,* 1988, 107; Soraya de Chadarevian, "Sequences, Conformation, Information: Biochemists and Molecular Biologists in the 1950s," 1996: 379 and ff.
[22] Joseph S. Fruton, "Early Theories of Protein Structure," *Annals of the New York Academy of Sciences* 325, no. 1 (1979), 1-20; Frederick Sanger, "Some Chemical Investigations on the Structure of Insulin," *Cold Spring Harbor Symposia on Quantitative Biology* 14(1950), 153-60; Miguel García-Sancho, "A New Insight into Sanger's Development of Sequencing: From Proteins to DNA, 1943–1977," 2010: 272 and ff.
[23] Werner Kogge, 2012: 609 and 31.
[24] James Watson and Francis Crick, 1953: 965 (see above).
[25] Soraya de Chadarevian, "Sequences, Conformation, Information: Biochemists and Molecular Biologists in the 1950s," 1996: 379 and ff; Miguel García-Sancho, "A New Insight into Sanger's Development of Sequencing: From Proteins to DNA, 1943–1977," 2010: 289 and ff. See also Soraya de Chadarevian, *Designs for Life: Molecular Biology after World War Ii* (Cambridge: Cambridge University Press, 2002), Part III.
[26] On Crick's interests in neuroscience, see Christine Aicardi, "Of the Helmholtz Club, South-Californian Seedbed for Visual and Cognitive Neuroscience, and Its Patron Francis Crick," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 45, no. 1 (2014), 1-11. On the discovery of transfer RNA, Hans-Jorg Rheinberger, *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube* (Stanford: Stanford University Press, 1997). On Sanger's RNA sequencing and other previous and contemporary efforts, Jérôme Pierrel, "An Rna Phage Lab: Ms2 in Walter Fiers' Laboratory of Molecular Biology in Ghent, from Genetic Code to Gene and Genome, 1963–1976," *Journal of the History of Biology* 45, no. 1 (2012), 109-38: 121 and ff. On the discovery of messenger RNA, and Nirenberg and Matthaei's approach to the genetic code, Lily Kay, 2000, 223 and ff.
[27] Sydney Brenner [1963], "Letter to Max Perutz," in *The Nematode Caenorhabditis Elegans,* ed. William B. Wood (New York: Cold Spring Harbor Laboratory, 1988), X-XI.
[28] Quote from Sydney Brenner, "The Genetics of Behaviour," *British Medical Bulletin* 29, no. 3 (1973), 269-71: 269. The experimental results were reported one year later in Sydney Brenner, "The Genetics of Caenorhabditis Elegans," *Genetics* 77, no. 1 (1974), 71-94. French molecular biologists François Jacob and Jacques Monod had proposed similar notions of information and program during their early 1960s experiments on genetic regulation in bacteria. At that time, Brenner had intensely cooperated with these researchers. See Hans-Jorg Rheinberger, "The Notions of Regulation, Information, and Language in the Writings of François Jacob," 2006; Lily Kay, 2000, ch5.
[29] On the connections between genetic and computer programs in Brenner's *C.elegans* research see Soraya de Chadarevian, "Of Worms and Programmes: Caenorhabditis Elegans and the Study of Development," 1998; Miguel García-Sancho, "From the Genetic to the Computer Program: The Historicity of 'Data' and 'Computation' in the Investigations on the Nematode Worm C. Elegans (1963–1998)," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43, no. 1 (2012), 16-28. On the more general transition from telegraphs to computers as technological models of gene action see Evelyn Fox Keller, 1995.
[30] Miguel García-Sancho, *Biology, Computing and the History of Molecular Sequencing: From Proteins to DNA (1945-2000)* (Basingstoke, UK: Palgrave-Macmillan, 2012), 49 and ff. especially 58-62. See also Frederick Sanger, "Sequences, Sequences, and Sequences," *Annual review of biochemistry* 57, no. 1 (1988), 1-29.
[31] James Watson and Francis Crick, 1953: 965; Francis Crick, Leslie Barnett, Sydney Brenner, and Richard J. Watts-Tobin, "General Nature of the Genetic Code for Proteins," 1961; Francis Crick, John S. Griffith, and Leslie Orgel, "Codes without Commas," 1957. In an influential historical essay written in the late 1960s, Gunther Stent distinguished between an "informational school" of molecular biology – focused on the mechanisms of gene action, including the coding problem – and a "structural school" – focused on X-ray analyses of the three-dimensional shape of biomolecules. None of these schools addressed the chemical sequence of DNA, RNA or proteins until Sanger's migration to the LMB. Gunther S Stent, "That Was the Molecular Biology That Was," *Science* 160, no. 3826 (1968), 390-95. Prior to Sanger's migration, there had been attempts to introduce amino acid sequencing into three-dimensional studies of the structure of proteins, but they had proved inconclusive. Soraya de Chadarevian, "Sequences, Conformation, Information: Biochemists and Molecular Biologists in the 1950s," 1996: 370 and ff.
[32] Frederick Sanger, "The Croonian Lecture: Nucleotide Sequences in DNA," *Proceedings of the Royal Society of London. Series B. Biological Sciences* 191, no. 1104 (1975), 317-33: 317.

[33] On the use of computers in Sanger's laboratory see Miguel García-Sancho, *Biology, Computing and the History of Molecular Sequencing: From Proteins to DNA (1945-2000)*, 2012, ch.3; Glyn Moody, *Digital Code of Life: How Bioinformatics Is Revolutionizing Science, Medicine and Business* (Hoboken, New Jersey: John Wiley and Sons, 2004), 15-16. On the introduction of minicomputers in biomedical laboratories, Joseph November, *Biomedical Computing: Digitizing Life in the United States* (Baltimore: Johns Hopkins University Press, 2012), ch.3; Bruno J Strasser, "Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954–1965," *Journal of the History of Biology* 43, no. 4 (2010), 623-60; Timothy Lenoir, "Shaping Biomedicine as an Information Science," in *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, ed. Mary Ellen Bowden, Trudy Bellardo Hahn, and Robert V. Williams (Medford, New Jersey: Information Today, 1999), 27-45; Edna Suárez-Díaz and Victor H Anaya-Munoz, "History, Objectivity, and the Construction of Molecular Phylogenies," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 39, no. 4 (2008), 451-68; Joel B Hagen, "The Introduction of Computers into Systematic Research in the United States During the 1960s," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 32, no. 2 (2001), 291-314; Hallam Stevens, *Life out of Sequence: A Data-Driven History of Bioinformatics* (Chicago: University of Chicago Press, 2013). On the features of minicomputers and the transition from mainframes, Paul E. Ceruzzi, *A History of Modern Computing* (Cambridge: MIT Press, 2003), ch.4; Martin Campbell-Kelly and William Aspray, *Computer: A History of the Information Machine* (Boulder, Colorado: Westview Press, 2004), Part Two.

[34] On the use of minicomputers as text processors, see Thomas Haigh, "Remembering the Office of the Future: The Origins of Word Processing and Office Automation," *Annals of the History of Computing, IEEE* 28, no. 4 (2006), 6-31. On the adoption of the algorithms – commands – of early text processors by the first DNA sequencing software, Miguel García-Sancho, *Biology, Computing and the History of Molecular Sequencing: From Proteins to DNA (1945-2000)*, 2012, 88 and ff. Fox Keller has shown that the mid-to-late 20th century witnessed a transition from the telegraph to the computer as the technology on which biological processes of living organisms – including gene action – could be modeled. In a previous paper, I argued that, within the category of 'the computer', the shift from mainframes to minicomputers was also crucial to the current understanding of the gene. Evelyn Fox Keller, 1995; Miguel García-Sancho, "From the Genetic to the Computer Program: The Historicity of 'Data' and 'Computation' in the Investigations on the Nematode Worm C. Elegans (1963–1998)," 2012: 20 and ff.

[35] Lenny Moss, 2004, ch. 1 especially pp. 44-50.

[36] Sydney Brenner, "The Genetics of Behaviour," 1973: 271; Sydney Brenner, "New Directions in Molecular Biology," *Nature* 248(1974), 785-87. There is a still unresolved debate among philosophers of biology on whether the original notion of information, as postulated by Crick in the central dogma of molecular biology, is compatible with the role of environmental, extra-genetic mediation, in organismic processes. See Paul E. Griffiths and Karola Stotz, *Genetics and Philosophy: An Introduction*, 2013, chs. 4-6; Karola Stotz, "Molecular Epigenesis: Distributed Specificity as a Break in the Central Dogma," *History and Philosophy of the Life Sciences* 28, no. 4 (2006), 533-48. Alex Rosenberg, "Is Epigenetic Inheritance a Counterexample to the Central Dogma?," *History and Philosophy of the Life Sciences* 28, no. 4 (2006), 549-66. Marcel Weber, "The Central Dogma as a Thesis of Causal Specificity," *History and Philosophy of the Life Sciences* 28, no. 4 (2006), 595-610; Kenneth Waters, "Causes That Make a Difference," *The Journal of Philosophy* 104, no. 11 (2007), 551-79.

[37] Frederick Sanger, "The Croonian Lecture: Nucleotide Sequences in DNA," 1975: 317, 18 and 28 and ff.

[38] The term repeatedly appears in the Croonian Lecture, as well as in Sanger's more technical articles on DNA sequencing, including those in which he presents the plus-and-minus and dideoxy methods: Frederick Sanger and Alan R Coulson, "A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase," *Journal of molecular biology* 94, no. 3 (1975), 441-48; Frederick Sanger, Steven Nicklen, and Alan R Coulson, "DNA Sequencing with Chain-Terminating Inhibitors," *Proceedings of the National Academy of Sciences* 74, no. 12 (1977), 5463-67.

[39] Oxford English Dictionary, online edition at http://www.oed.com/ (last accessed May 2014).

[40] Allan M Maxam and Walter Gilbert, "A New Method for Sequencing DNA," *Proceedings of the National Academy of Sciences* 74, no. 2 (1977), 560-64; Allan M Maxam and Walter Gilbert, "Sequencing End-Labeled DNA with Base-Specific Chemical Cleavages," *Methods in enzymology* 65, no. 1 (1980), 499.

[41] Walter Gilbert, "DNA Sequencing and Gene Structure," in *Nobel Lectures, Chemistry 1971-1980*, ed. Tore Frangsmyr and Sture Forsen (Singapore: World Scientific Publishing, 1980), 408-26: 408 and 24. Contrast with use of 'read off' in Sanger's lecture, reprinted at Frederick Sanger, "Determination of Nucleotide Sequences in DNA," in *Nobel Lectures, Chemistry 1971-1980*, ed. Tore Frangsmyr and Sture Forsen (Singapore: World Scientific Publishing, 1980), 431-47.

[42] Victor A. McKusick and Frank H. Ruddle, "A New Discipline, a New Name, a New Journal," *Genomics* 1, no. 1 (1987), 1-2: 1. See also Alexander Powell, Maureen A O Malley, S Muller-Wille, Jane Calvert, and John Dupré, "Disciplinary Baptisms: A Comparison of the Naming Stories of Genetics, Molecular Biology, Genomics, and Systems Biology," *History and Philosophy of the Life Sciences* 29, no. 1 (2007), 5-32: 7 and ff.
[43] Walter Gilbert, "A Vision of the Grail," in *The Code of Codes: Scientific and Social Issues in the Human Genome Project*, ed. Daniel J. Kevles and Leroy Hood (Cambridge: Harvard University Press, 1992), 83–97: 92. Leroy Hood, "Biology and Medicine in the Twenty-First Century," in *The Code of Codes: Scientific and Social Issues in the Human Genome Project*, ed. Daniel J. Kevles and Leroy Hood (Cambridge: Harvard University Press, 1992), 136-63: 158.
[44] Lenny Moss, 2004, 44; Lenny Moss, "The Meanings of the Gene and the Future of the Phenotype," *Genomics, Society and Policy* 4, no. 1 (2008), 38-57: 45.
[45] Contemporary gene-centric discourses have acquired the form of genetic determinism and/or reductionism. Genetic determinism is the belief that genes are the only determinant of organismic features. Reductionism is a close variant which was deployed by molecular biologists and states that from the structure of DNA one can determine how genes work. See A. Tauber and S. Sarkar, 1992; Sahotra Sarkar, *Genetics and Reductionism* (Cambridge: Cambridge University Press, 1998).
[46] See transcript of the ceremony in which the first draft of the human genome sequence was presented at http://clinton5.nara.gov/WH/New/html/genome-20000626.html (op cit. 1, last accessed June 2014). At this ceremony, the expression 'book of life' was used by the director of the HGP, Francis Collins. However, this expression possesses a much longer history and can be traced back to the origins of Christian thought. Molecular biologist Robert Sinsheimer explicitly referred to the human genome as the 'book of life' in 1967: Lily Kay, 2000, 30 and ff; Robert L. Sinsheimer, *The Book of Life* (Reading, Massachusetts: Addison-Wesley, 1967).
[47] Research on the history and philosophy of science has traditionally approached concepts as both epistemological and socio-political entities. My story suggests that there is also a technical dimension of concepts which should be taken into account when investigating them. On the connections between concepts and artifacts, and how concepts enable to do things beyond abstract thinking see Fernando Broncano, "Movilidad De Conceptos Y Artefactos," (Madrid: Universidad Carlos III, 2009, unpublished paper available at http://portal.uc3m.es/portal/page/portal/grupos_investigacion/hermes/Fernando_Broncano_Rodriguez/Movilidaddeconceptosyartefactos2.pdf); Mieke Bal, *Travelling Concepts in the Humanities: A Rough Guide* (Toronto: University of Toronto Press, 2002), Introduction and ch.1. On debates around the concept of gene and proposals of new notions, such as 'system', 'switch' or 'battery' see Evelyn Fox Keller, *The Century of the Gene* (Cambridge: Harvard University Press, 2000); Paul E. Griffiths and Karola Stotz, "Genes in the Postgenomic Era," *Theoretical Medicine and Bioethics* 27, no. 6 (2006), 499-521; Jane Calvert and Joan H Fujimura, "Calculating Life? Duelling Discourses in Interdisciplinary Systems Biology," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 42, no. 2 (2011), 155-63; Vivette Garcia and Edna Suárez-Díaz, "Switches and Batteries: Two Models of Gene Regulation and a Note on the Historiography of 20th Century Biology," in *The Hereditary Hourglass. Genetics and Epigenetics, 1868–2000*, ed. Ana Barahona and Hans-Jorg Rheinberger (Berlin: Max Planck Institute for the History of Science, Preprint number 392, 2010), 59-84.