A MATTER OF STYLE. HOW MAP THINKING AND BIO-ONTOLOGIES SHAPE CONTEMPORARY MOLECULAR RESEARCH

by **Federico Boem**

PhD Thesis in
Foundations of the Life Sciences and Their Ethical Consequences (FOLSATEC)

Submitted on

Supervisors:

Prof. **Giovanni Boniolo**
Dipartimento di Scienze della Salute University of Milano & Department of Experimental Oncology, European Institute of Oncology (IEO)

Dr. **Paola Roncaglia**
EBI-EMBL, Cambridge, UK

Lab Supervisor:

Dr. **Gioacchino Natoli**
Department of Experimental Oncology European Institute of Oncology (IEO)

European School of Molecular Medicine (SEMM) and University of Milan

Acknowledgements

First, I want to thank all my colleagues at FOLSATEC programme. More than fellows, these people, who I have now the privilege to call 'friends', have definitely contributed to increase and shape my current knowledge not only in an academic sense. In particular I want to thank Emanuele Ratti, my desk-mate and a real comrade in this venture of philosophically exploring the contemporary face of molecular biology. All the discussions we had and the conferences we attended together were always great moments of intellectual growth. Another great source of inspiration was Pierre-Luc Germain. Pierre Luc was both a colleague, an informal supervisor and a model to look at. By combining both great philosophical knowledge with technical and scientific expertise, Pierre Luc was an indispensable source for my study. Moreover, his humility and generosity were also a model to tend to for being a true intellectual. A special mention goes also to my colleagues in the lab. Especially Pietro Lo Riso, Giulia Barbagiovanni and Giuseppe D'agostino, who helped me through out these four years by introducing me to the most sophisticated details of molecular biology. Of course I want also to thank Giovanni Boniolo. He was not just a supervisor, he was a mentor who never stopped to encourage me both with his comments and critiques. Another big thank goes to Paola Roncaglia, my external supervisor who helped me a lot with the technical part on Gene Ontology and with the entire revision of my work. In the same way a special mention goes also to Gioacchino Natoli. His sharpness and rigour, signs of a true scientist, were an example to my work as well. I cannot forget also David Teira, who was fundamental in shaping the structure of the thesis and as a source of very interesting material. Last but

not least, the biggest thank goes to my parents. I owe them everything, from genetics to culture. I would not be here without their encouragement and their love.

ABSTRACT

The aim of this thesis is to provide an epistemic analysis of the transformations occurring in contemporary biological research by considering the relation between molecular biology and computational biology. In particular, I will focus on *bio-ontologies*, as the tool which incarnates at best the new face of biomedical research. Such a choice is not arbitrary. By appealing to the notion of *style of reasoning* and *way of knowing*, I will show that bio-ontologies exemplify the rise and success of *map thinking* as the signature of a new way of doing molecular biology, while the theoretical tenets, established more than 30 years ago, still maintain their epistemic prominence. This is neither to say that experimentalism will disappear from science, nor that the experiments power will be diminished but rather that experiments will have a new role in the architecture of scientific efforts, precisely because of the increasing importance of classificatory approaches. Therefore, such a transition within biomedical research is indeed radical and profound but it does not involve paradigm shifts but rather a change in the practice. In this sense, it is a matter of style.

Outline

CHAPTER I

In this chapter I provide an epistemic reconstruction of the rise of classification strategies in biomedical research. In particular I analyse this phenomenon through the lens of the map, meaning that most of the projects of Big Science are aimed at building "maps". I analyse the notion of map, its features and its relation to models. In particular I sketch how maps can constitute tools of surrogate reasoning. Then I show how map is both a well known metaphor in biology but also represents a specific style of thinking. Thus I show how map/model are connected with map reasoning.

CHAPTER II

In this chapter I analyse the peculiar epistemic status of biology in terms of laws and theory. I analyse the reasons behind biology's epistemic disunity, its virtues and problems. Then I examine both historical roots and epistemic motivation for such a picture. In the end I show how the need for unity has been invoked by many researchers in the life sciences and how this fact promoted the implementation of bio-ontologies.

CHAPTER III

In this chapter, first I distinguish philosophical ontology from computational ontologies, highlighting differences and connections. Next, I explore the rise and the motivation of bio-ontologies as a response to the need of unification in biology. After a general description I focus on GO, the most famous tool of this kind in biomedical research, by showing its virtues and limits. Second I describe the relation between ontologies and databases as a connection among different

types of maps. In this context I specify why GO, by providing a semantic synthesis of experimental results, is an orienteering tool for biomedical research. Moreover, I also describe how ontologies, born as a descriptive tool, became also normative and in which sense they are so. I defend the idea that these classificatory tools, neither entirely a theory nor precisely a model, constitute a new epistemic category. Ontologies are in a sense a surrogate of theory in biology as they unify but in a bottom-up fashion rather than imposed from above.

CHAPTER IV

In this chapter I examine some, selected publications heavily based on bio-ontologies, to provide an empirical grounding for the increasing success and implementation of bio-ontologies in biomedical research. In doing so, I would insist on the novelty of these approaches in terms of style, by showing that such articles would have not been published 15 years ago. By showing that the molecular view still holds, despite the radical change in methodologies, these examples show that the actual transformation in biology is indeed a matter of style (of reasoning) rather than a paradigm shift.

CHAPTER V

In this chapter I will briefly examine the phases of development of molecular biology in order to show how map thinking, begun with the Human Genome Project, constitutes a new phase of the research. I explain that, contrary to common interpretation, bioinformatics should not bee seen as a new discipline besides molecular biology but rather as a new moment of molecular biology development. I show how such a picture involves several levels of analysis, often intertwined, from epistemic reasons to social institutionalisation of scientific disciplines. Next I frame this in terms of style of reasoning by showing how map

thinking penetrated all the life sciences, thus reshaping the directions of future research. Last, I analyse the challenges, the risks and the possible implications of such a new order of things.

INTRODUCTION

In recent years biomedical research has changed. It is still changing. This should not surprise anyone. History of science is full of transformations. In the last thirty years biology has been shaped by the triumph and penetration of experimental methods. The last thirty years were indeed the age of molecular biology. The *molecular turn* redefined and modified the approaches, the scope, the practice of the life sciences at any level, from cell biology to ecology. Now it seems that we are facing a new venture. It is what we may call the *computational turn*. Again this does not simply mean that biologists moved from the benches to the computer screens. What has changed and it is still changing is also, and more importantly, the way researchers think, how they prove what they claim, how they justify their results. From small laboratories biology went Big.

Interestingly, the computational turn has not just transformed the technological apparatus. Neither simply the methodology. Computational methods have surely speeded up the power of analysis and granted higher and higher volume of data be examined. However, a more profound transition is at stake. An epistemic one. Indeed the very objects, the data that scientists have to

deal with have changed. Cell culture is not just what biologists can observe and manipulate in their Petri dishes any longer. It is also a codified line within a database. Experimental findings are nowadays classified in larger and larger electronic repositories. The meaning of local experiments is now directly confronted and compared with results coming from other labs. Moreover, databases are not just keeping data. They are ordering, classifying, structuring data. They establish the scientific meaning of these data. Databases are then shaping biological knowledge.

One may ask what the nature of such mutations is. Is it just a question of different models and their application? What does it mean that scientists have changed their way of thinking? If it is true that classifications, databases, collection strategies, bio-ontologies are all changing the face of science, then the very aim of this thesis is to provide a philosophical analysis of how and why it is so.

On methodology: a matter of style

Contemporary scientists do not usually debate about the nature of their explanations. A group of researchers may dispute on what model should be adopted to answer a certain problem, or whether an experimental strategy will be either promising or not in order to achieve a particular result. But scientists will not address the question about what a model is. Neither they will discuss why and how experiments are the right tool to build a scientific evidence. Traditionally, this is a task for philosophers. The question then, is how they do so. Unlike scientists, philosophers do not have tangible tools, specific materials,

experimental apparatuses. Yet they possess methodologies too. Philosophical instruments are conceptual. This does not mean they are not technical. The ambition of philosophy of science is precisely this: to deal with science through a language that might resemble the ordinary one but is indeed technical and yet aims at avoiding triviality and oversimplification.

One may see scientific explanation as a particular way to justify certain beliefs about nature. However, there is not just one. In mathematics, if we have to justify why we hold that Pythagoras' theorem is valid for all the right-angle triangles, we usually appeal to the notion of *formal proof*. Yet geometrical demonstration is a mode of justification that is very different from empirical confirmation. Nevertheless we consider these diverse modes, according to diverse contexts, perfectly suitable. Thus one may wonder why such a difference actually exists.

Philosophers have elaborated several conceptual categories to describe and represent scientific narratives and cultures. One may see scientific transitions as clashes either of different *models* or *theories*. For instance, the phenomenon of combustion can be seen as the passage from a theory employing certain entities (*i.e.* phlogiston) to another one framed in modern chemistry (thus something 'burns', among other conditions, due to the presence of oxygen). This is more or less what W.V.O. Quine (1960) had in mind with his notion of *conceptual scheme*. Different schemes may be 'logically' compared so that, allow me to simplify, one is true while the other is false. Thus scientific facts are here naively, directly, comparable.

In contrast, Thomas Kuhn (1962) famously proposed that the very same models and theories might be embedded in diverse confronting *paradigms*.

Models within the same paradigm can be compared. However between paradigms there is a sort of *incommensurability*. This is due to the fact that scientific terms, despite they might have the same name, mean different things in different paradigms. Kuhn indeed showed that a comparison between scientific theories cannot be reduced to a simple truth-values assignments of the sentences held by those theories. Going further, Paul Feyerabend (1975) notoriously highlighted the extra-logical, irrational, elements of scientific discovery and justification. At first glance one may build a hierarchy of these categories by arguing that models are constructed within different paradigms. In this sort of matryoshka game, Crombie (1994) and Hacking (1985, 1994, 2004, 2012)[1] formulated the further category of *styles of reasoning*, or styles of *scientific thinking*. Accordingly, disparate styles arose in different periods of human history, and they provide distinct systematic approaches to deal with the real world. Compared to other analytical tools, the introduction of the notion of style constitutes a novelty both in history and philosophy of science. Indeed it is a way to analyse scientific changes through history from a perspective that is exquisitely epistemic. It is also an approach to deal with scientific theories and claims that is closer to cultural comparative anthropology. Styles, since they are overarching categories providing what may count as an explanation, also set the problem of scientific objectivity to another level of epistemic investigation. Indeed *objectivity* itself is the result of a particular style of reasoning[2]. This is because it is always the style, through self-authentication, that allows to frame its

[1] Hacking's view, although stemming from Crombie, is not entirely reducible to Crombie's. However for our purposes here it is not important, at the moment, to highlight such distinctions.

[2] See for instance, Daston and Galison 2007 and Hacking 2009

specific rules of truthfulness. Hacking adopts *truthfulness*[3] instead of truth because he argues that styles' epistemic peculiarity rests not just on what facts are represented either by true of false statements, but rather on how (according to which way of thinking) such a conclusion about their truth or falsehood is reached. As Crombie writes styles "introduced new objects of scientific inquiry and explanation, new types of evidence, and new criteria determining what counted as the solution of a problem" (1994, vol.1, p 83). According to both authors it seems it is not clear how to select and establish the number of styles.

Following Rasmus Winther (2012a), who systematises both Hacking and Crombie, styles of reasoning can be grouped and listed as following: *axiomatic*, *experimental*, *hypothetical/analogical*, *taxonomic*, *probabilistic* and *historical/genealogical*. Let us briefly examine each of them.

- The axiomatic style, formulated in ancient Greece, deploys the use of formal proofs. Although the link between logic and other areas of mathematics (*i.e.* geometry) has been rationally reconstructed only *a posteriori* after Frege's turn, both disciplines rest on the power of formal demonstrations.

- The experimental style questions nature through observation and measurements and also by constructing artificial devices in order to elicit natural phenomena.

- The hypothetical/analogical style adopts theoretical idealisation to uncover real properties of the world.

- The taxonomic style makes distinctions in terms of hierarchies and similarities.

[3] Following Bernard William's distinction (2004)

- The probabilistic style provides a decision criterion when facing uncertainty.
- Finally, historical/genealogical style gives reason to natural phenomena by building their historical roots and development.

Surely, such a list is not exhaustive but it definitely grasps the most important ways of thinking in the history of science. Each style can also be associated with iconic figures (*e.g.* Galileo, through his mathematisation of nature, as the exemplar of the hypothetical/analogical style) or objects (*e.g.* Boyle's air pump for the experimental style), which show how the style crystallises a specific way of thinking and doing.

However it should be clear that these analytic categories such as models, theories, paradigms and styles are not mutually exclusive. On the contrary, as argued by Winther (2012a), although certainly distinct, they are definitely mutually intertwined. For any category runs the risk to impose a unique, or at least, privileged, epistemological account of what science does, both in theory and practice. Indeed science is complex because the world is so. I agree with Winther (2012a) that a proficuous approach would be to adopt more than one category at the same time. The task is ambitious but has the merit to elicit aspects of scientific cultures that, otherwise, would be neglected. Following Hacking suggestions, Winther examines the different dynamics of these "interweaving categories". Such an interaction involves multiple levels of perspective and different types of relations. The first is the so called "realization relation", affecting models, paradigms and styles so that "the latter member […] instantiate and implement the former member" (Winther, 2012a, p 632) thus meaning that the "three categories are nested in an abstraction hierarchy" (*ibid*). In contrast

"the three categories exist on the same level of abstraction" (*ibid*). The "guidance relation" describes how "higher level categories constrain the properties and parts of lower level categories" (*ibid*). In the end, "inheritance relation" highlights how properties and parts of 'objects' at a higher level can be transmitted to lower levels. This does not mean that an analysis based on interweaving dynamics would be just combining a particular style with that peculiar paradigm which in turn shows certain models. Forms of combination can occur also within such categories. As Winther puts it "[t]he actual working of science include multiple realization among category levels and hybridization within a level" (*ibid*). Accordingly, he provides, as an example, an analysis of biological systematics by using different "interweaving categories". To briefly sum up Winther's work, he reconstructed (2012a) the development of naturalistic classification by showing how such one discipline presented different *styles of reasoning* (naturalists surely began using taxonomic and genealogical styles but nowadays molecular mechanisms regarding phenomena such as gene duplication are considered fundamental, as well as mathematical modelling, for any scientific phylogenetic enterprise). The very same discipline oscillated among, at least, three *paradigms* (from a Linnaean perspective, via an evolutionary approach, to algorithmic framework of gene distribution frequencies) and employed different *scientific representations* (either mathematical ones or metaphors) of the structure of the living beings (from the chain of being to the tree of life, to the network of life).

My idea is to provide a similar philosophical analysis, framed into a general scheme that takes into accounts the dynamics of diverse interweaving categories,

for the current situation in molecular life sciences, after what I called the *computational turn*.

CHAPTER I

From molecular biology to bioinformatics: the map thinking

A shift?

In 2008 the famous American magazine *Wired* had on its special issue's cover a provocative title: "The End of Science". Chris Anderson, the former editor in chief of *Wired*, explained that sentence by arguing, more in details, about the "end of theory" in science. According to Anderson the image of scientific disciplines, still guided by theoretical hypotheses, should be considered obsolete, and thus be abandoned in favour of a new picture. This means that the new face of science will be shaped by different approaches, new ways of doing. More precisely, such a novelty should be understood bearing in mind the new challenge provided by so called Big Data Science. The label of Big Data does not mean just a big volume of data[4]. Despite the lack of a precise definition, it is certainly possible to select certain features of Big Data Science as they were 'hallmarks' for such an approach. Following Kitchin (2013, 2014) Big Data Science consists certainly in the quantity of data (*e.g.* petabytes), in the speed at which these data are obtained, in the variety in which they are ordered and displayed, in the global/holistic aim (in contrast to more traditional statistics), in standardised procedures both regarding resolution and identification, in their relational format which can be easily expanded or increased in magnitude. Examples of *big data projects* are now easy to find within the context of

[4] If just the amount of data counts, also taxonomy and astronomy could be seen as Big Data Science. Other features then seem to be required to establish what contemporary scientists mean by Big Data Science.

biomedical research. Let us mention some of them. The ENCODE project, aimed at providing a comprehensive map of all regulatory elements of the human genome, is certainly one of the most famous (and controversial one, see for instance Germain *et al.* 2014). Another good case is represented by the NIH Roadmap Epigenomics Project, funded "with the goal of producing a public resource of human epigenomic data to catalyze basic biology and disease-oriented research. The Consortium leverages experimental pipelines built around *next-generation sequencing* technologies to *map* DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts in stem cells and primary ex vivo tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease" (http://www.roadmapepigenomics.org/, emphasis is mine). Again, there is the 1000 Genomes Project, whose purpose is to sequence the genomes of a large number of people in order to provide a more exhaustive map of human genetic variations. As written on its webpage, "[t]he goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations studied. This goal can be attained by sequencing many individuals" (http://www.1000genomes.org/about).

Of course, Big Data Science is not feasible without computer science. In the last years, due to the development of powerful computational tools, many studies in the life sciences are now possible just because of a discipline called *bioinformatics*. As a matter of fact, defining what bioinformatics is, is not an easy task. As for many disciplines (for instance, where the boundary between biochemistry and molecular biology actually lies?) bioinformatics is a combination of different methodologies. For our purpose it could be sufficient to

claim that bioinformatics "is conceptualizing biology in terms of macromolecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied maths, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale" (Luscombe, Greenbaum and Gerstein 2001).

The rise of such a field should not be read as meaning that a change in the practice of biological research can be explained due to the mere introduction of computers in biological studies, beside and beyond the bench work. As a matter of fact, computers have been part of biological research since years. It is rather the practice according to which computational resources are used and how they foster a transformation in the way research is done and justified, that makes the real difference.

This is not just an epistemic claim. The wind of change is perceived and promoted also at more mundane level. Since the last decade more and more funding agencies have supported the so called *data-driven* approaches which have been sold (certainly, also for economic purposes) as capable to revolutionise the entire scientific enterprise by reforming the nature of science itself. *Data-driven* is a term used to designate a particular way of doing scientific research. Accordingly, science should move from hypotheses testing, lab exploratory experimentations and theories formulation, to the direct address of raw data. Data should primarily guide research. However the picture portrayed by media might appear quite fuzzy. Despite the ease of their use (also by professional scientists), *data-driven* and *Big Data Science* are not interchangeable. Data-driven approaches denote a methodological stance that

privileges data collection and pattern recognition over theoretical hypotheses formulation. This does not mean that data-driven cannot be implemented in a small lab. On the other hand, the description of such techniques seems to perfectly suit large scale experimental endeavours suggesting how data-driven would be tailored for Big Science. Thus it is the combination of these two aspects that is often invoked as the key feature of a new scientific era.

Indeed, several publications (*e.g.* Mayer-Schönberger and Cukier, 2012) have supported the argument that in contemporary research, the very idea about the centrality of hypotheses will be replaced by a new way of doing, towards a perspective completely centred into the analysis of patterns coming from pure data. Such a popularised view describes this turn as just the emergence of a new scientific endeavour meaning that, after the rise of molecular biology in the 1960s and 1970s (see for instance Morange 2000, 2006) which implemented *experimentalism*[5] within the life sciences, now a new transformation comes from a supposed computational revolution.

The philosophy within

However, despite the abundance of reviews and descriptions, few attempts have been made to provide a philosophical analysis for such a change. Setting the epistemic primacy of concepts and technologies regarding these phenomena is not an easy task. Once one enters the historical path of bioinformatics, the first impression is to face a vicious circle. On one hand, the rapid development of computational instruments allowed the analysis of larger and larger datasets. On

[5] Roughly speaking, the idea that 'truths about nature' can be discovered and justified through specific tests conducted under particular and controlled conditions (for a panoramic overview concerning the discussion of the experimental dimension in biology see, among the others, Mayr 1982, Rheinberger 1997, Weber 2005)

the other hand it is precisely the increasing relevance of large databases that pursued the need for more adequate instruments.

This might suggest why Big Data and data-driven, even if they do not mean the same thing, are often coupled together. If it is true that science is going Big, data-driven are invoked as the right tool to deal with it. Next Generation Sequencing (NGS) technologies are now capable of accessing information coming from whole genomes in a single run and are able to critically analyse these data to infer potential features of genome's behaviour. The power of these tools opened the doors to the so-called *data deluge* or the fact that the amount of data produced oversteps the possibility of their analysis. However, some scholars (*e.g.* Strasser 2008, 2012a, 2012b) argued that *data-driven* is not a novelty in the history of biology. Natural history also relied on collections and data comparison to build scientific claims. Thus the *computational turn* in the life sciences, if we want to call it so, it is certainly in the methods employed, but also in the kind of technological devices adopted and, more importantly, in the justification strategies provided for such a practice.

However, the current status of biological research cannot be described simply as the revival of classificatory reasoning. Indeed, if it is true that genomic libraries of DNA fragments can be seen as a molecular version of naturalistic collections, contemporary electronic databases, unlike material collections, allow faster and more precise procedures in gathering, organising, using and re-using data. Moreover these collections are still embedded in the conceptual framework that has shaped biology since now, namely *molecular revolution*. Regardless of its historical novelty or not, such an information explosion could be also a potential harm for research as it could, both theoretically and practically, impede

the accessibility to data. Thus, data curation and organisation became an indispensable and yet ordinary task for current science. Such an effort is definitely eased by informatics tools as it has to be standardised and uniform for the entire scientific enterprise. In other words, once different databases are settled, then a sort of meta-database (a structure to order, compare and integrate information coming from diverse databases) is required. One approach in this direction is constituted by semantic instruments, developed in the field of information management, namely *applied ontology*. It should not be a surprise then, to ascertain how this field of research is getting more and more relevance. Indeed if someone searched for the term 'ontology' on Google he/she could be surprised to realize that the first entries mainly refer to applied ontology. In this battle for notoriety, 'ontology' in a more traditional and philosophical sense is defended just by Wikipedia and a few of other websites. While *philosophical ontology* was devoted to pure speculation, engineers and computer scientists revitalized such a notion in the light of its possible applications. Indeed, in modern computational jargon a *computational ontology* is a way to model and represent a domain of interest or a particular area of knowledge so that a computer can process it. As Gruber pointed out (Gruber 2009) "an ontology specifies a vocabulary with which to make assertions, which may be inputs or outputs of knowledge agents (such as a software program)". Here lies the difference. If *philosophical ontology* was pursued as a way to establish on pure speculative ground 'what there is' or the fundamental entities or things of the world, *applied ontology* is a subfield of informational research devoted to knowledge representation and data integration. To make a slogan from 'ontology' we came to 'ontologies'. Again, as Gruber writes "ontologies are

typically specified in languages that allow abstraction away from data structures and implementation strategies; in practice, the languages of ontologies are closer in expressive power to first-order logic than languages used to model databases" (Gruber 2009). In other words, ontologies constitute a tool that allows comparison among data that were originally produced and stored in different manners. In addition, ontologies are conceived as the mode to translate a specific knowledge at a certain level of description to other levels. This is why ontologies are also said to be the "semantic level" of scientific modelling.

Biomedical research is one of the leading areas of inquiry for the implementation and application of these semantic instruments. *Bio-ontologies* (as they are called) are now proliferating in the management of many biological databases. Among them, the *Gene Ontology* (from now on GO), developed by the Gene Ontology Consortium, represents a promising project greatly employed by many different institutions and laboratories in all the life sciences. The semantic dimension of this enterprise is clear in its own mission. The aim of the Gene Ontology project is to provide a representation of the features of gene products across different species and databases through a controlled vocabulary of different "biological categories".

Nevertheless, the centrality of databases for contemporary biological research does not rest on pure technological innovation. Indeed, putting data together, just in quicker or more efficient ways, cannot constitute, *per se*, a genuine form of conceptual change. If, following Strasser, we can agree about the existence of a *fil rouge* connecting natural history with contemporary biology, then we have also to specify in what do they differ from each other. Surely the implementation of specific computational tools in biology cannot be

described as just a technological advance. Thus, if *collection* (e.g. see the Harvard Museum of Natural History or the Natural History Museum in London) has been a prominent style of reasoning in natural history, this should be examined also in the cultural context of its practice and in relation with the relative theoretical paradigms. Understanding electronic databases, or tools like Gene Ontology, requires that contemporary collection strategies must be set in the new framework of molecular biology. Here lies the limit of the adoption of a single analytic category representing scientific cultures. Instead, given the peculiar epistemic status of that framework (see for instance Rheinberger 1997), a combination of *interweaving categories* seems to be unavoidable.

The map thinking

In this chapter I try to frame all these categories around the map notion. The idea is to characterise the epistemic culture of contemporary biomedical research in relation to the idea of the map. Indeed, maps can stand for particular kinds of scientific models (see further), they can also be seen as paradigmatic conceptual metaphors driving the research (*e.g.* the search for the map of DNA regulatory elements in order to better understand genome's behaviour), and they can be thought as a particular way of doing, namely the map thinking as the regulatory ideal of how scientific research should be pursued. All these different levels must be kept together.

In this section I will discuss what a map is, what are its features and limitations and why its adoption is central for the understanding of the practice of contemporary biological research. My argument is that the notion of map, as a key concept of current research, serves, at least, two purposes. First, it debunks

the idea that scientific efforts can be clearly distinguished in terms of hypotheses versus data. Map thinking does not ban hypotheses from science, it rather changes the relations between data gathering and hypotheses formulation. Second, map thinking embeds a specific way of reasoning which has a long, although often neglected, tradition in the life sciences. A philosophical and historical analysis of the notion of map would help to better frame and comprehend the practice of contemporary research.

Despite the efforts of depicting science as a unified, yet articulated, enterprise by prominent members of the scientific community, the rise of computational *data-driven* approaches fostered heated discussions about the nature of research itself, the way science is defined (Weinberg 2010, Golub 2010, Brenner 2010). These discussions, wittingly or not, revealed a deep disagreement concerning the epistemic hierarchies within scientific methodologies, thus regarding hypotheses generation, experimental strategy and design versus data production and collection. If, on the one hand, big data projects (*e.g.* the ENCODE project[6]) have promised to change the face of research, on the other hand more traditional biologists contested the very theoretical foundations and the methodological approaches of such findings. If it is true that the mechanistic understanding of traditional molecular biology is still the unavoidable mode of explanation to claim the presence of causal connections, it is also true that new methodologies allow a "30,000-foot view" (Vogelstein *et al*. 2013) to highlight connections and relations which are not detectable from "the ground".

[6] The aim of ENCODE is to provide a comprehensive *map* of DNA's functional elements

Nevertheless, such discussions have also highlighted a possible change in the aims of the research itself. The need for and the creation of *biological maps* (the map of the genes, the map of the proteins, etc.) is often described as the primary interest of contemporary biologists. However, this way of doing science is sometimes accused (*e.g.* de Chadarevian 2009) to miss the very nature of scientific endeavour. According to some detractors, global maps are surely more comprehensive than traditional approaches of molecular biology and certainly they provide insights on global behaviours, but they lack understanding of biological mechanisms, they lack *causality*. The tone of the controversy has been sometimes vitriolic (see for instance Graur *et al.* 2013 in which scientists involved in the ENCODE project are defined as "genomic clochards") revealing somehow that the question is not just about technological advance, but more profoundly about conceptual changes in scientific research.

Despite the emphasis, I argue that the core of discordance is not in construction of biological maps. The notion of map is, on the contrary, what actually unifies these perspectives. It is not the map itself, it is how scientists build (and should build) the map and what do they think counts as a map. It is the style of reasoning. In a recent interview about Big Data in molecular research, Sydney Brenner precisely worries that "[n]obody understands what proof is in biology" (de Chadarevian 2009, p.68). Brenner thinks that the way certain claims are supported in that type of research is not justified. More precisely that way does not count as a legitimate justification. Here Brenner is involuntarily adopting Hacking and Crombie epistemic tools, as he is practically saying that it is indeed a matter of style of reasoning. Again, I think that the important

differences here lie not in *what* scientists are doing but rather in *how* they are doing it.

Let us consider the endeavour of understanding the genome's behaviour. Molecular biologists too used the expression "mapping" to explain the meaning of their efforts. Since an overall approach for a global comprehension was not feasible and probably not even conceivable, the main strategy was to reconstruct it in an additive stepwise fashion. Ancient explorers did not have satellites. They constructed their maps piece by piece.

The problem thus is not the map, the metaphor/model adopted by both molecular and computational biology, but rather the way map reasoning is justified. The change, in my opinion, lies at a different level, in what counts as a map and what are the right modes to support certain claim on and of a biological map. Thus the epistemological challenge here is then to explain how and why a distinct way of reasoning is so problematic (or at least, why it is perceived so) and nevertheless rapidly succeeding. To do so, it is arguable also to follow Rheinberger's recommendation in defining epistemology: "the concept is used here […] for reflecting on the historical conditions under which, and the means with which, things are made into objects of knowledge. It focuses thus on the process of generating scientific knowledge and the ways in which it is initiated and maintained" (Rheinberger 2010 p 2-3).

In reconstructing the history of the rise of this aspect of computational biology my focus will be on *map building*. Surely one may ascribe maps to the *taxonomic style* of reasoning. Maps are ordering, under a common dispositional rule (i.e. a form of classification), dispersed and fragmented knowledge. But then it is fundamental to analyse how this style embeds a different *Weltanschauung* in

diverse historical moments of the life sciences. I mean that even if the map still holds as a successful picture representing past and present-day scientific efforts in the field of biological research, the notion of map itself, of what a map is, has instead changed. Epistemic intersections happen diversely at different levels. Naturalists were surely mapping things as contemporary biologists are doing now. However computational classifications rest on a theoretical framework that has more in common with molecular biology than with natural history. If the molecular revolution constitutes a paradigm shift, this does not mean that previous styles of reasoning have been replaced. On the other hand, the nature of such a scientific enterprise and its focus on classification and database construction cannot be simply explained as a paradigm shift. These new technologies have not shaken down the great overarching assumption that led scientific development to the, so-called, *molecular revolution*. Following Morange, it seems that the depiction of such novelties in terms of a transition "from a reductionist to a holistic vision of biological phenomena" (Morange, 2006. p 23) is too simplistic. Rather than rejecting the molecular perspective, these new approaches seem to actually reinforce it but putting the emphasis on a "different level of organization" (*ibid*.).

Let us consider briefly the case of genomics. Genomics (the global study of the structure and the behaviour of the genome) has not made genetics obsolete, neither the theoretical assumptions of the first are discordant or in conflict with the ones of the latter. Again the passage from traditional molecular studies to computational approaches is not so linear, neither explicable just by an opposition of contrasting theoretical schemes. Therefore, leaving aside oversimplified reconstructions, the building of an epistemic framework for these

innovations should take into account "that data-driven science seeks to hold to the tenets of the scientific method, but is more open to using a hybrid combination of abductive, inductive and deductive approaches to advance the understanding of a phenomenon" (Kitchin 2014). Again, a single category seems being reductive. Indeed other scholars claimed that the real innovation of these approaches does not rest on a new capacity for providing explanations but rather on a diverse, and more structured way to generate hypotheses (see for instance Ratti 2015).

To sum up, the tension between these approaches is definitely more complex. Let us just consider the case of the Human Genome Project (HGP). The HGP is an international research project aimed to establish the DNA sequence of the entire human genome. "HGP researchers deciphered the human genome using three tools: producing what are called linkage *maps*, complex versions of the type originated in early Drosophila research, through which inherited traits (such as those for genetic disease) can be tracked over generations; making *maps* that show the locations of genes for major sections of all our chromosomes; and determining the order, or "sequence," of all the bases in our genome's DNA" (https://www.genome.gov/11511417, emphasis is mine). Therefore, one may rightfully claim that HGP constitutes the exemplar of map thinking. However the HGP, unlike current projects, has not been pursued, albeit it required a profound computational power, through new Next Generation approaches. Indeed it was certainly a map. It seems definitely to constitute a conceptual 'rupture', the first example of a new way of thinking but, given the technology through which it has been conducted, it is also the link between traditional molecular biology and current research.

In order to better grasp this point, let us briefly focus on two recent examples. These two cases are settled at two different levels of research. The first case, a project highly framed on contemporary efforts towards translational research and personalised medicine, is represented by a recent venture developed by the University of California that, by combining patient information, clinical data and scientific findings, aims at building the first "Google Map for Health" (Leuty 2015). The emphasis on map reflects the idea that deeper comprehension of complex biological phenomena (in this case also involving social factors) can be better and more genuinely achieved through the construction of maps. The second case comes directly from cancer research. The study addresses the problem that biological knowledge coming from molecular biology is *dispersed* (I will focus on this aspect in second chapter). Indeed scientists argue that [m]ost existing pathway databases provide a view on molecular mechanisms as disconnected processes, splitting their content into 'canonical pathway' representations. However, cancerogenesis affects simultaneously multiple cellular processes and their crosstalk. Therefore, cancer research can benefit from the reconstruction of cellular signalling in the form of a *comprehensive map*, representing the complexity of pathway crosstalk manifested by co-participation, interaction or co-regulation of molecular entities in several cell signalling processes. Understanding connections between molecular mechanisms is important for determining potential therapeutic intervention points" (Kuperstein et al. 2015, p 1, emphasis is mine). As a result, researchers have built an interactive and dynamic database, the *Atlas of Cancer Signalling Network* (ACSN)[7] which provides (also in a graphically fashion display) the entire current

[7] Note the semantic choice of 'atlas' that clearly refers to the geographical dimension

map of signalling processes in non cancerous cells but often disrupted during tumorigenesis. Such an effort shows very well the virtues ascribed to maps in developing biological research. First, the ACSN is the most comprehensive map of cancer mechanisms (due to the fact that it is database is based on the most recent literature that can be easily updated). Second, it presents such information through a graphic display that can be browsed via *Google Maps itself*. Third, there is a related discussion forum in which scientists can debate results and configuration display. Last, from the map it is possible to directly access the primary information coming from other related databases.

Arguably, map thinking is a way of reasoning that overarches different technological approaches and it is not reducible to any of them. It is also a mode to do science that does not represent a precise theoretical framework as a Kuhnian paradigm. As a matter of fact, the practice of building maps in the life sciences is older than computational biology. I suppose it is a matter of style. A style of reasoning wavering between different paradigms, employing several models.

Maps as styles, models and images

By following Hacking's scheme, mapping activity can be definitely linked to taxonomic style. Natural history begun with collecting and classifying. According to Pickstone (2001 p.60) the expression *natural history* should be intended as the "register of facts" of the natural world. Thus in contrast with *natural philosophy* which was rather involved in the searching for an explanation for those facts. Two different styles of reasoning indeed.

Taxonomic style is not just listing what is out there. Collecting things presupposes a collection design, meaning it requires to set up what these collected 'things' are. It also involves comparison criteria, what should count as similarity, where to put boundaries. Indeed classifying does not mean just 'grouping together'. Unlike experimentation, whereas claims and results are local, collection shows a broader epistemic horizon. The laboratory of collectors and classifiers is the world itself. The analogy with geography is not accidental.

The word 'map' comes from the Latin *mappa,* which seemed designating a napkin on which maps were drawn. Yet etymology is not of great help. In the Renaissance 'map' was the short version of *mappaemundi* or 'map of the world'. Thus a map is a graphical representation of geographical areas. It is then an image, a description. In a figurative sense a map is a detailed representation. Moreover, a map is an ordered representation. It is so because it reveals the, supposed, hidden order of things. Maps represent relations and connections virtually invisible to empirical observation. In turn, representation is neither reducible to depiction nor comparable to simple description. Representations work due to their *idealized* content. World's facts are complex and noisy. Many factors are at play at the same time. It is very hard to discriminate among all these components just by looking at them. Maps, by exalting certain features rather than others, provide an *idealized* situation in which important elements and traits are revealed. A map serves its task when it allows someone to orientate himself/herself. A map as big and detailed as the land it represents would have an undesirable feature: its uselessness. On the contrary, maps are effective precisely because they display some kind of information and neglect the other. Of course, different maps fulfil different purposes and thus select different information to be

shown. The map of *The London Underground* is not respectful of the distances between stations. However it perfectly satisfies its scope, that is to guide Londoners and millions of tourist navigating in one of the biggest metropolis of the world without paralysing it. As Jacob argues "a map has the power to create in a given space […] 'window' that opens onto another space, suited to a form of intellectual mastery different from everyday empirical perception" (Jacob 2006, p.99). Moreover a map constitutes a very peculiar type of representation. A map stands for what it represents, but also involves 'creation' of something that is not present in its referent and allows theoretical conjecture, expectations, predictions. Again "[…] the map is no longer simply a record of cumulative knowledge. It is also a heuristic mechanism that lends itself to interrogation, hypotheses, a quest for explication, and the testing of numerous correlations between human phenomena and natural settings" (Jacob 2006, p.370).

In this perspective a map acts as a *model*. Models constitute a hot topic in both science and philosophy. To sum up, contrary to the Syntactic View, that privileges logical axiomatisation as the main grounding of theories, the so-called Semantic View (see for instance Van Fraassen 1980) gives more credit to models, even suggesting that theories are sets of models. Moreover, from considering just mathematical models, philosophers and scientists now acknowledge that models constitute a plurality of 'entities'. Beside mathematical models (as the over-quoted Lotka-Volterra model of predation) now we can count concrete models (as the DNA scale model of Watson and Crick), pictorial schemes (as pathways representations in molecular biology), computational models that simulate the behaviour of a phenomenon by focusing on the actors of the process depicted and their interactions. The list can be easily extended. This

attention to modelling as the core of the scientific practice, had deep consequences on epistemology. Thus the focus of philosophers has gradually moved from formal consistency to empirical adequacy meaning that the relevant thing to explain is the determination of the type of relation intercurring between the world and the model itself. However there is nothing new under the sun. How models stand for real things, or how do they accurately describe phenomena, is just a restating of the problem of *similarity* (for a more detailed discussion see for instance Suàrez 2003, Weisberg 2013), which is central to philosophical investigation since Plato. Among different philosophical accounts, there is, to a greater or lesser extent, a consensus view on the fact that models represent aspects of the world. The question then is what the meaning of 'representation' is. Such a topic is directly intertwined with the problem of scientific realism. The debate is then endless and ramified in more or less sophisticated positions (see for instance Giere 2006, Fraassen 2008). It is not of my interest to address this problem here. It is more relevant to me trying to examine how scientists use models and why, rather than tangling up with the formulation of a universal account of scientific representation. Indeed, after the 'pragmatic turn'(see for instance Hacking 1983, Cartwright 1983), many philosophers have started to concentrate their attention on the practice of science, meaning that a more precise analysis of what scientific research is, should start from a comprehension and description of scientists' activities. In other words understanding science should begin by looking at what scientists do rather than establishing what science should be. According to this perspective, the question about models turns on what models grant and allow. Thus "[t]o explain a phenomenon is to find a model that fits into the basic framework of the theory and that thus allows us to

derive analogues for the messy and complicated phenomenological laws which are true of it. The models serve a variety of purposes, and individual models are to be judged according to how well they serve the purpose at hand" (Cartwright, 1983 p 152). In sum, the fact the models represent, means that they are a *device* through which one may learn about the *object* represented. In other words, as the model *stands* for the object (or the process) represented, it is possible to understand some properties of the object itself precisely by acting on the model. Models allow then forms of *surrogative reasoning* (Swoyer 1991). Surrogative reasoning means that building and studying a model allows scientists to reason about features regarding the system the model stands for. Scientists can learn, make hypotheses and design experimental strategies through models. In addition, a model is more than a visual representation (meant as a detailed picture) since it is a dynamical entity. The very process of model building and the constant interaction between the scientist and the model are two key components of this type of surrogative reasoning. It is the *manipulation* of the model (Morgan 1999), the possibility to intervene in the model, that allows researchers to make predictions and to devise possible explanations about phenomena. Precisely according to such a *pragmatic view* of models (Winther 2012b), maps can be definitely seen as a particular type of models. As for other scientific idealizations, map's generalisation acquires a meaning in a given context and with given purposes.

Like models, maps should be *used* in order to be understood. Unlike certain kinds of models such as tables and schemes, maps have distinctive formal constraints. First, maps are highly dynamical. Second, maps are not simply depictive, they rather share common characteristics with so called *constructional*

drawings, that are graphic representations of the constructing procedures in the design of a project (see for instance Maynard 2005). Moreover, as proposed by Valeria Giardino (Giardino 2013), maps can be classified as particular kind of models, specifically concerned with a type of representation that privileges spatial correspondence and structural features. Indeed, among models, maps have peculiar construction rules that favour measuring procedures. It is in their structural properties that also lies the understanding of what types of manipulations are allowed or not. Contrary to models of mechanisms, a map normally does not involve the possibility of adding/removing elements in order to observe alternative configurations. It rather grants to select a precise point or an area on itself and to examine its relations and dispositions with the remaining parts. In this sense maps provide a sort of standardisation of representation by creating a common, shared frame. As reported by Winther (2014) a map can perform distinct forms of generalisation/idealization. A map can simplify the messiness of the represented target, aggregate different elements under the same label, exaggerate features in order to highlight them, enhance some details over others, displace or hide certain sorts of information, create types by constructing relations and connections etc. As argued by Winther (2014) these idealisation procedures often result in the production/discovery of certain kinds. However this must not be taken as a strong metaphysical turn. Such kinds are deeply epistemic. "Geographic features, processes and objects are of course real. Yet, we must structure them in our data models and, subsequently, select and transform them in our maps" (Winther 2014 p.15). These kinds are thus the result of an idealization. However, this again does not mean that they are imposed a *priori*. One can just start building a map from (supposedly) unrelated information

and then observe connections and patterns ‘emerge’. Moreover such a mapping activity does not always require a graphical display.

In this sense maps are not just physical objects. Mapping is definitely an activity that refers to a way of doing. Indeed, in terms of knowledge production and intervention mapping gives also a tremendous power. In his masterpiece *Guns, Germs and Steel* (1999) Jared Diamond wonders how Francisco Pizarro and a few other *Conquistadores* were able to conquer the Inca Empire and defeat its 80.000 soldiers. More provocatively, one may ask why Incas did not try to conquer Spain. Of course Europeans had a better technology (firearms, armours) and invisible allies (germs of diseases unknown to native Americans) but their success, according to Diamond, depended on another fundamental factor. “A related factor bringing Spaniards to Peru was the existence of writing. Spain possessed it, while the Inca Empire did not. Information could be spread far more widely, more accurately, and in more detail by writing” (Diamond, 1999 p.78). In other words, writing allowed the Conquistadores to *map their knowledge* against the magnitude of natural phenomena. No one could have all the possible experiences about the world in his or her life. However one could read and learn from books. Writing is a way to map knowledge coming from different sources in order to make it reusable. A map then is a tool to observe and represent certain features of the world without the need to start every time from scratch. Thus, if maps are a useful way to find directions, to orientate in the complexity, they provide also an image to deal with such a complexity. Maps have a metaphorical power too.

The map metaphor is not new in biology. It is one of the famous images of nature along with the *chain of being* and the *tree of life* (see Barsanti 1992,

2005). The father of modern taxonomy, Linnaeus, by criticizing the older image of *Scala Naturae* (literally, the 'ladder of nature'), a mode of classification considered too linear, in his *Philosophia botanica* writes that living beings "dispose themselves as the Territory on a Geographical Map" (Linnaeus 1751, p 77, English translation is mine). The representation of Life was changing. Natural affinities were recognized as too complex to be portrayed linearly. That is why Linnaeus had in mind that species should have been represented as circles, blending into one another. Paul Dietrich, one of Linnaeus pupils, actually drew that map. The map was then not just a theoretical assumption in Linnaeus' mind. It was a real picture of the natural world. And the new image of nature was efficacious as it precisely responded to a new comprehension of nature itself. As Barsanti writes, unlike the *Scala*, "the Map presents […] a two dimension territory that is virtually possible to go through in any direction and from which […] it does not emerge any tendency" (Barsanti, 1992, p 50, English translation is mine).

However, in this battle of images, the map progressively lost ground in favour of the *tree of life*. Evolutionary thought imposed a different style of reasoning. Living beings have a history that should be reconstructed genealogically. Again, another style of reasoning. This turn has not been painless. Such an change still has consequences on debates about natural classification[8]. More interestingly, the rise of experimentalism condemned mapping activities to oblivion. Biology, after the molecular turn, is now a science as physics, not stamp collecting any longer.

[8] see for instance *In Defense of Classification* by John Dupré, 2001

Map thinking regained consideration with the Human Genome Project. Before the HGP, molecular biologists, like sailors and explorers discovering specific geographical *loci*, have studied single genes and molecular circuits. With the HGP these genomic *loci* needed to be considered together. It was the time for "mapping the code" (Davis 1991). Next Generation Sequencing was far to be achieved. However this is the period in which databases started to gain the more and more influence. Nowadays no-one could think to do research in biology without relying on at least one database. This 'trend' is perfectly represented by the increasing importance of the *Omics*. Such a suffix nowadays usually refers to a set of disciplines sharing both epistemic and technological similarities. Omics are mainly aimed at the global analysis of genes, gene expressions, proteins, metabolites, biologically relevant interactions in a given sample. Thus from genomics (aimed at providing a comprehensive list of genes), a term invented in the late 80s just before the design of the Human Genome Project, now we have transcriptomics (profiling mRNA) and epigenomics (looking at global epigenetic modifications), proteomics (proteins), pharmacogenomics (at the intersection between genomics and pharmacology) and interactomics (the set of molecular interactions). The list is not exhaustive indeed. From a technical point of view, all the Omics adopt so called *high-throughput* (HT) screening techniques to generate large amounts of data, which are in turn thought as fundamental to permit a system-level comprehensions of interactions and relations between different elements. The classificatory reasoning is at the foundations of this way of doing. "In high throughput research, knowledge discovery starts by collecting, selecting and cleaning the data in order to fill a database" (Schneider and

Orchard 2011). Once again from the time of natural history mapping is part of scientific endeavours.

Some prominent scientists have argued (*e.g.* Gilbert 1991) that, in the years of the HGP, biology was undergoing a paradigm shift. The terminology, as I tried to previously defend, in my opinion is not correct. Surely, as I have shown before, biology changed, but the centrality of the molecules (*e.g.* DNA, RNA, proteins) as the right level of both investigation and explanation was not contested by the new approach. Interestingly the mere implementation of techniques does not constitute *per se* an unavoidable move towards data-driven. Maps are not enough to call for the end of hypotheses. Biology is a combination of styles. Contemporary biological maps are also and still embedded in an experimental way of doing. In confirmation of this consider that Gilbert also claimed that such a turn would have transformed molecular biology, from a set of techniques, to a more intellectual enterprise. "[A]ll the genes will be known (in the sense of being resident in databases available electronically), and that the starting point of investigation will be theoretical. An individual scientist will begin with a theoretical conjecture, only then turning to experiment to follow or test that hypothesis" (Gilbert, 1991, p 99). This is precisely the opposite of what the advocates of computational biology claim nowadays and what makes people like Sydney Brenner upset.

Map thinking and contemporary biology

Certainly, as for natural collections of the past, the building and the use of maps put scientists into a different *level of abstraction* (the wrong one according to Brenner) which is not the single cell examined in a lab. Indeed the horizon of

possible conjectures and inferences is broadened. This is because the map reasoning allows to count as a fact what was hidden before or even unconceivable. In molecular research an experimental result can be grounded on the consistency of the methods adopted and on the locality of its production, namely, the *experimental conditions*. As also remarked by François Jacob "in biology, any study [...] begins with the choice of a 'system'. Everything depends on this choice: the range within which the experimenter can move, the character of the questions he is able to ask, and often also the answers he can give" (Jacob 1988). Thus biological findings seem to be strictly dependent on the locality of their production. Contemporary biological maps were built and thought to overcome such a locality. Let us consider, as an example, the case of ENCODE.

ENCODE's aim is to provide an encyclopaedia of DNA elements, such as transcripts (coding or not), binding sites, enhancers, insulators of the human genome. ENCODE can be seen as the further step of the Human Genome Project in terms of both techniques and mapping efforts. However the two projects rest on slightly different intellectual accounts concerning the nature of the genome. Indeed, before the HGP the view among the majority of biologists was that once the complete 'code' had been 'cracked', this would have granted a global understanding of all relevant biological phenomena. Lewontin (2001) has nicely shown how the language adopted to explain and promote the HGP was full of colourful expressions, even towards sort of mystical tones, regarding the challenge of revealing the "secret of life".

Nowadays the genome is pictured differently. The completion of the genomic map, rather than a conclusion, fostered entire new areas of investigation and provoked profound discussions. The image of the genome itself has changed.

The complexity of the genome requires not just mapping 'objects' as *loci*, but also relations, *e.g.* how these elements behave and interact. I would argue then that it is the new way of doing, that introduces new things and categories, that can, and should, be mapped. The molecular paradigm is still there, slightly modified by the increasing relevance (now also 'molecularised') of non genetic factors. The map metaphor holds. But the style seems to have radically changed. The key word here is *regulation*. ENCODE is thought of as a map of DNA regulatory elements. This does not mean that regulation is something new (even within the *genocentric* framework scientists know that genes are regulated somehow). It rather means that regulation is the cornerstone of the new way of doing research. It is what to look at. From being marginal, regulation became central. Such a change is not caused by ENCODE, it precedes it.

This transformation of perspective should be intended as framed into the so-called *epigenetics revolution* (Carey 2013). A unique definition of what epigenetics is still lacking. Some people may refer to DNA methylation, others may extend it to the environmental contribution to genome's behaviour, others again, by following Waddington as the "whole complex of developmental processes that connects genotype and phenotype" (Waddington 1942, 2012 *reprinted*). As a working definition I can stipulate that epigenetics is the study of those non-genetic factors that regulate gene expression and that can be preserved/transmitted through both mitosis and meiosis. However, I will not delve into the complexities required for disentangling the semantic ambiguity traditionally attached to 'epigenetics' (see for instance Jablonka and Lamb 2014). Sometimes, in science, concepts work precisely because of their fuzziness (see for instance Rheinberger 1997). I rather want to consider the epigenetic

contribution in terms of styles of reasoning. In other words, I am interested here in examining what epigenetics pushed scientists to do, consider and experiment on. The rising importance of epigenetics (Heijmans and Mill 2011, Mill and Heijmans 2013, Landecker and Panofsky 2013, Meloni and Testa 2014) redefined what to look at in many areas of biological research. As a result, even if the exact relevance of epigenetic phenomena is still debated and a consensus view on the contribution of such phenomena compared with genetic ones lacks yet, few scientists disregard epigenetics as such. ENCODE is definitely a project plunged into a different understanding and image of the genome.

Surely, ENCODE efforts are, like the HGP, on mapping. But, in some respect, ENCODE is a different map. First, it reflects the theoretical move from a structural/topological conception of genes to a functional one (Gerstein *et al.* 2007, Stamatoyannopoulos 2012). It also promoted a heated debate on what a function is in biology (see Germain *et al.* 2014). Second, it completely embeds a new way of addressing and representing what is considered important. If we want to adopt Brenner's words, ENCODE is mapping another *level of abstraction*. This also explains why people like Brenner think it is the wrong one. Such a map, according to them, portrays simply the wrong way to reason, the wrong things to look at. Other scholars (Graur *et al*. 2013) went further and despised the entire project by claiming that maps like that are not science, since they are just piling data. In my opinion this is also because people like Graur are, more or less consciously, defending an image of biological research as anchored to and defined by a precise style of reasoning. Accordingly, a map like that is not science, it is stamp collecting. On the contrary, as I recently wrote with two colleagues, "[j]ust like 'collections' maps are different from a simple

accumulation of data: they are rather ways of organizing data according to specific aims, in order to make specific contrasts emerge and to enable specific kinds of investigations. As we have argued […], the mapping of the biological activity of DNA elements produced by the ENCODE project is indeed a 'collection' in that sense, that allows biologists to generate precise hypotheses about the biological role of certain DNA elements" (Germain *et al*. 2014, p 828).

A consequence of such a different way of doing is mirrored also by changes in the social practices of science. Indeed, it is fundamental to note that such 'mapping' could not be achieved by a single lab. These efforts must come from network activities thus suggesting another analogy with natural history. Natural history has always been a collective enterprise. Müller-Wille (2007) has shown how, at the time of Linnaeus, the correspondence between botanists and naturalists was thick and widespread. Ariane Droescher (2008, p 152) argues that when Linnaeus in 1742 received a sample of *Linaria* (a genus of plants commonly known as toadflax) which showed a radial symmetry instead of bilateral one, he deduced in 1744 that he had a specimen of a new species and not simply a new variety. Linnaeus was able to do all of this by considering just one sample and without conducting any kind of experiment. Linnaeus did not make experiments as he was adopting a different style of reasoning. His results were possible because of the power of naturalistic collections. More importantly, he did not base himself just on his own data. The horizon of his investigation has been broadened by the capacity of referring to other collections. Indeed Emma Spary (Strasser 2006, p 113) has defined natural history as "a science of networks". Considering the appropriate, historical differences, systematics was a form of Big Science too. The potential of this science of networks should be

intended as both at practical and theoretical level. From a practical point of view, such a cooperative research allows technical possibilities that a single lab or group could not achieve (in terms of available instruments, raw material and financial support). From a theoretical perspective, this capacity of going 'global', beyond the locality of the single experimental context, is precisely what allows forms of unification and generalisations that a single lab or group could not accomplish. Thus "[…] ENCODE is an instance of 'big science', involving 442 members from 32 institutes and a budget around 288 million USD. In late 2012, this effort led to the simultaneous publication of 30 papers in *Nature*, *Genome Biology* and *Genome Research*. […] ENCODE is also a massive amount of publicly available data, with a total of 1,649 high-density experiments on 147 cell types (at the time of the 2012 round of publications). The project's contributions also include technical standards (both in 'wet' protocols and computational analysis), novel tools or algorithms, and a careful assessment of the strengths and weaknesses of different technologies" (Germain *et al*. 2014, p 809).

Yet ENCODE only refers to the human genome. It says nothing about other organisms. It is not its fault. The human genome is its scope. Moreover ENCODE purposes are mainly directed to biomedical applications. ENCODE has not been thought to unify biological knowledge or perspectives. Surely, ENCODE puts together insights coming from various sources, but it does not build a unified picture of biological phenomena. ENCODE is a map but not of that kind. Thus, albeit comprehensive in its context and scope, ENCODE is still just a tiny drop in the ocean.

This situation is not fortuitous. From a physicist's perspective, biology is a science *sui generis* . Unlike physics, biology does not present a clear, well defined theoretical unity. This is due also to the absence of general laws and principia. Besides, despite the praise for a unified experimental approach, many areas of biology developed independently from each other by adopting different models, jargons, vocabulary and implementing diverse methodological protocols and standards. In this sense maps proliferation did not help. On the contrary, the vast majority of biological databases are built in different manners and the information provided cannot be easily compared or/and integrated. It is biological knowledge here that has to be mapped.

Indeed, biological knowledge looks like a Babel Tower. "For the most part, the current systems of nomenclature for genes and their products remain divergent even when the experts appreciate the underlying similarities. Interoperability of genomic databases is limited by this lack of progress" (Ashburner *et al*. 2000). Biology needs unification. This is what bio-ontologies are thought for.

To sum up, in this chapter I argued that contemporary biomedical research is shifting towards the map thinking (the return of collection strategies beside the experimental practice), picturing an idea of knowledge grounded on the image of the map (biological knowledge should look for more generality) and producing many maps of data (*e.g*. ENCODE project, Cancer Genome Atlas, etc.). In doing so, I have also specified how collecting is a genuine (and not entirely new) form of scientific enterprise along with experimentalism and how maps can be considered as a legitimate type of models. In the end, I have anticipated that all these efforts try to cope with the fact that biological knowledge is epistemically

dispersed. The image of biology as a Babel Tower is more than a joke. It is a real problem. In the next chapter I will try to provide an answer for such a situation (*i.e.* the, *sui generis*, epistemic status of the life sciences) and I will describe a possible solution both from a practical and theoretical point of view.

CHAPTER II

Biology: its epistemic status and its disciplinary unity

Biology began its history struggling a lot to be allowed among the other sciences. The shame of just being stamp collecting loomed over it for decades. Nevertheless it became one the of most proficuous ones, and nowadays it has probably the strongest and most significant implications on human life, culture and society. The epistemic status of biology itself, its autonomy within sciences, its disciplinary boundaries, are all problematic. Edmund Wilson (1901, p 19), reports that, in the late 19th Century, those who nowadays are named *biologists* were rather professional figures performing at least three different 'jobs'. First there were *bug-hunters* or naturalists. Naturalists were collecting, classifying and comparing. The second job was performed by *worm-slicers*, mostly physicians, working on comparative anatomy and morphology. In the end there were *egg-shakers*, interested in cell physiology applying methodologies imported from chemistry and physics. Three different styles of reasoning indeed. Biology, as we know it today, has yet to come. According to Giulio Barsanti (2005) these three 'styles' developed independently one from one another. Anew, in the late 19th Century Michael Foster (1899) sadly reports that anatomists, zoologists and physiologists behave like strangers and, even if they would like to communicate to each other, their disciplines are not mutually intelligible. Indeed a clash of styles. Hacking has argued (1985) that sometimes, once a new style replaces another, it makes it very difficult for scientists accustomed to the new style to recognise the previous objects of investigation.

Such a fragmentation was still present in the first half of the 20th Century. Natural history and systematics, embedded into a Darwinian framework and thus adopting a genealogical style, stood up against Mendelian genetics that took advantage of a combination of experimental and statistical way of doing (see for example Barsanti 2005). The rise of Modern Synthesis solved some theoretical issues but was still far from producing a unified framework for all the life sciences. Biology has still many 'mothers'.

Indeed these styles kept pervading biology through time. Sometimes with strong conflicts. It is the case, for instance, of the so called "molecular wars" (see Wilson 1994) whereas the rise of molecular biology in the 1960s and 1970s has been perceived as a direct attack to evolutionary biology and natural history. Mayr's (1961) famous distinction between *functional* and *evolutionary* biology, was also aimed to restate the autonomy of certain ways of doing research in the life science, in respect to molecular studies. Later on, despite the final and decisive imposition of a molecular perspective in many areas of the life sciences, biology was yet to be unified. Cytology and classical genetics survived to the molecular turn and it is important to remember the 1980s were definitely the "golden age of cell biology" (Morange 2000, p 244). History of science is never plain and linear. History of biology, due to its multiple origins and its ramified structure, is even more complicated and intricate.

As also claimed by Morange, the 'victory' of molecular biology (see the previous chapter) can be seen as the ascendancy of a certain type of methodological reductionism. However such a form of methodological unification based on reduction was limited given that "the recourse to higher levels of analysis is utterly indispensable if the edifice of modern biology is to

remain intact. Only such a reference to higher levels will enable the molecular biologist (and the biochemist) to understand the finality of the biological phenomena they study, and thus to justify their research" (Morange 2000, p 246). Thus even if the molecular turn has probably imposed a more unified perspective on the way of doing, a strong theoretical unification still lacks. The many 'mothers' of biological research are also present within the molecular view. This is perhaps because the objects of biological investigation are many and they are very different from each other. The diversity of the living beings is one of the primary interests for biologists. Biology, among other sciences, has the merit to have put the focus on what is *distinct* rather than looking just at what is *similar*. This does not mean that this epistemic focus is the right one. Again, I suppose, it is a matter of style. Following Wimsatt (2007), the ontology of such a science is more a "tropical rainforest" than a Quinean desert.

Biology, theory and laws of Nature

The peculiar epistemic status of biology is shown also by the issues regarding its theoretical dimension. A common, folk view is that scientific truths are represented and exemplified by certain laws. The idea that scientific truths might be discovered and faithfully represented through *laws of nature* is a notion emerged in the time of the so-called Scientific Revolution (see for instance Rossi 1997, Giere 1999). Laws of nature were initially originated as a curious blending of theology (if there are laws there must be a legislator) and mathematics (reflecting Galileo's claim that the 'book of Nature' is written in a mathematical language). Following a secularised version of Descartes and Newton ideas, 20th Century science generally has conceived laws of nature as *true generalisations*

about world's phenomena which do not depend on space and time. In addition, they are also thought to be *necessary*, to distinguish them from contingent regularities. Newton's Laws of Motion can be considered a paradigmatic example of this kind. However, a more precise investigation shows that such laws "[…] are neither universal nor necessary – they are not even true" (Giere 1999, p90). No thing in the universe precisely obeys to such laws. Natural laws present a high degree of *idealization*. Surprisingly, they grasp aspect of the world by distorting it. Indeed representing does not always mean picturing a faithful description. Van Fraassen (2008) has nicely shown how distortion played a central role in the effectiveness of pictorial perspective during the Renaissance. However, when we say that a representation is *effective*, it is fundamental to clarify what does it mean. Efficacy can be spelled out in different terms. As Weisberg (2007, 2009) recalls, the activity of theorising displays always different epistemic goals or "representational ideals". Descriptive accuracy, completeness, logical consistency, causal explanation can be all considered a not exhaustive list of goals for a theory. These aims might not be achieved all together at the same time and "trade-offs" (Weisberg 2007, 2009) are unavoidable. For example, according to Cartwright (1983), natural laws have certainly a solid explanatory power about reality, even if they do not describe reality at all. According to many accounts, *laws of nature* illustrate then ideal situations that do not correspond to any actual case but they rather generalise and abstract the common features of those different situations in order to construct a uniform type of explanation. The literature on what exactly laws should grasp about natural world is endless and often intertwined with the debate about realism and metaphysics of science (see above all Cartwright 1983, Carroll 1994,

Van Fraassen 1989). However, I am not interested here in the ontological debate about the nature of laws. Thus, whether laws exist, if they are discovered or invented, it is secondary here. I am more concerned (as for models) with the fact that in scientific practice, researchers (chiefly physicists in this case) use laws as a formalism to make sense of the raw material of their investigation. The most important point is to recognise how such laws have been seen as constituting the theoretical core of a scientific discipline (namely physics) under which and according to which new observations, experiments and explanation must fit.

By this pragmatic perspective, at first glance, compared to physics biology seems very different. For instance "looking at biology […] one of the first things people notice is that there is apparently not much role for scientific laws" (Godfrey-Smith 2014, p11). Although there have been some disputes on the opportunity to describe Hardy-Weinberg principle in population genetics as a law (Sober 1993, Elgin 2003), it is also evident that it could be, at least in theory, thought to be reduced to its basic physical laws. Thus the question over biological laws is deeply intertwined with the problem of reductionism. In this sense one may also read Ernst Mayr's (2004) famous defence of the epistemic autonomy of biology from physics. Moreover, while some efforts have been done in order to determine whether there are laws in a particular biological subfield (*e.g.* population genetics or ecology), fundamental laws unifying all biological realm seem to lack yet. In addition, most of the debate has been influenced by the treatment that logical empiricism has given to the notion of law, and its connection with an *a priori* status, sometimes very distant from everyday scientific practice. Indeed, the absence of clear biological laws can be maybe attributed, again, to the differences among practices in the life sciences.

First of all, despite all the philosophical debates, biological regularities seem to lack the feature of *necessity* (in a sort of Leibnizean sense[9]) which is often invoked for the fundamental laws of physics. The suggestive (albeit epistemically naïve) picture is that Newton's space is empty and universal, while Darwin's world is blooming, full of exceptions, and teeming with different creatures. Physics is universal (in the sense that its laws are valid anywhere in the universe) and necessary, biology local and contingent (however on the role of contingency of biological phenomena see Gould 1996, 2002). Although both philosophically and scientifically imprecise and inaccurate, such view has historical reasons that can be found in the different roots of the scientific practices. Again, following Pickstone (2001), while *philosophia naturalis* (*i.e.* physics) looked at 'how the world goes', *historia naturalis* (*i.e.* biology) was concerned on 'what is there in the world'. As argued before, these different styles of reasoning shaped also the epistemic goals of these disciplines thus promoting one aspect over the other (*e.g.* the top-down search for laws over the bottom-up quest for classificatory schemes).

Second, the introduction of mathematical formalism, despite its growth in the recent years due to the development of systems biology and its efforts on producing a 'theory for biology', has not been really directed to the discovery of 'laws' but rather on the building of models. This aspect must be framed in the specific logic of discovery of molecular biology. Indeed, as famously argued by Machamer, Craver and Darden (2000) the fundamental aim of molecular biology is to discover *mechanisms* (and thus the models of them). The meaning of the term 'mechanism' in biology is not easily assessed. Craver and Darden define

[9] In contemporary debates such a position is advocated, in different ways, by so called *necessitarians* (see for instance Swoyer 1982 and Bird 2007)

mechanisms as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (2013, p 26). Such a definition reflects the philosophical enterprise of providing a general account of mechanism for biological practice thus aiming at unifying biology under the search for mechanisms. However science is not always more precise than philosophy and often scientific terminology is (intentionally or not) vague. A general account is far from being defined. Daniel Nicholson (2011) has shown how 'mechanism' can be spelled out differently in biology, waving from philosophical stances about the nature of living matter to the account for a particular type of molecular explanations[10]. The roots of such a situation are not simple to explain. Surely, the rise of the notion of mechanism in the life sciences has a long and complex history that cannot be addressed here. Yet, it is certainly true that, despite the semantic fuzziness of the term, mechanisms are, depending on the context, often invoked by contemporary biologists as way to certify their distance from several forms of *vitalism* and at the same time they are usually mentioned as the main *explanatory tool* of molecular biology[11]. Whether mechanisms actually stand for something real or they just constitute a profitable heuristic tool, again it is not strictly relevant to our discourse. It is not of my interest here to deal with the debate about mechanisms, their historical roots, what are they and how do they work. The central aspect to me is mainly on the epistemic use of mechanisms. It is a matter of style concerning the scientific practice. For instance, in the aims and scopes of one of the most important journals in the molecular field, *Molecular Cell*, it is clearly stated that the "[t]he

[10] This last one is the sense according to which Craver and Darden argue in favour of when they claim that biologists are above all concerned with mechanisms.

[11] Albeit different, these attitudes reveal a general will of biologists to put themselves close to physics. On the relation between physics and biology in relation to mechanisms see also Boniolo 2013.

journal focuses on analyses at the molecular level, with an emphasis on new *mechanistic insights*" (http://www.cell.com/molecular-cell/aims emphasis is mine). Molecular biologists are seeking for mechanisms not for laws.

Let us briefly consider an example of a paper by Robert Weinberg lab recently published on *Cell* (Chaffer *et al.* 2013) that can be genuinely considered as a paragon article for molecular biology. Leaving aside technical details, the article aims to show that the way a particular gene is expressed enables to explain a 'switch' in the behavioural state of certain groups of cells in breast cancer. The article is a very well reasoned set of precise investigations aimed at testing all the conditions of interest (*e.g.* what happens when a gene is knocked down or the identification of distinct cell populations through molecular markers etc.) within the experimental system. From a methodological point of view an article like that is often seen as exemplar by molecular biologists. But let us give a look at its conclusions. The authors write that "[t]he present work reveals that the dynamics of interconversion between epithelial non-CSC and mesenchymal/CSC states are important determinants of normal and neoplastic epithelial tissue behavior" (Chaffer *et al.* 2013 p 71). In other words, the researchers are saying that, in those specific experimental conditions, they were able to *observe* (an observation, *nota bene*, made possible by the experimental setting) a 'wavering conduct' or 'a back and forward transition', between groups of cells affecting the behaviour of the epithelial tissue. Then they claim to have shown that such a conversion of state in specific cells is *demonstrated* (obviously not in a logical sense) to occur "in certain carcinoma subtypes – notably, basal carcinoma of the breast" (*ibid*). These results are presented not just contradicting any supposed biological law but rather as contrasting a widely accepted model of

conversion of state which was mainly unidirectional. The 'solution' shown is indeed a mechanism. Finally the authors state that "this plasticity is not a *universal* property of all breast carcinomas" (*ibid*, emphasis is mine). The mechanistic model presented in the paper is not generalised nor precisely generalisable. As recently argued "there is no molecular biology paper that offers a description of a mechanism in terms of all of them. Actually, each paper is devoted to answer a different subset of questions, and such subset depends on the authors' interests and on the issue faced" (Boniolo 2013, p 264). Thus the virtue and the value of such a model of a mechanism are not in the provision of a ground for the construction of an overarching, formal, scheme which should encompass all possible instances, but rather in the fact that a mechanism constitutes a practical guide to be adopted for further investigations. These new inquiries will probably not use the mechanism as it is. Researchers will adapt and modify it according to the new experimental system, possibly adding some parts or removing others.

The emphasis on mechanisms over laws reveals also a lot about the problematic role of theory in biology. The search for mechanisms can be seen as unifying but they are more providing a collection of different techniques than a general conceptual groundwork. What is a theory in biology? Certainly, the Darwinian account of evolution and especially the so-called Modern Synthesis are often addressed as a "theory" but not really in the sense adopted for other scientific theoretical frameworks[12]. Evolution is definitely central to many biological efforts but it constitutes more a common stance or a background assumption than an overarching theory for biological world. Dobzhansky

[12] Mayr (2004) even argues that the theory of evolution is in reality a conflation of five different and epistemically distinct theoretical claims

famously stated that “nothing in biology makes sense except in the light of evolution” (1964, p 449) but many areas of research in the life sciences were pursued and still are with not direct remark on evolution itself (see for instance Germain *et al*. 2014). Thus when biologists and philosophers started to think about the role of theories in their work they first looked at what a theory is in other sciences.

Moreover, in biology, an attitude against ‘theory’ in favour of ‘experiments’ has a long and widespread tradition (see Lewontin 1970 and Callebaut 2013). Indeed, 20th Century biologists generally neglected the impact of conceptual work to their discipline and forgot how much it shaped the very beginning of the life sciences (*e.g.* consider the debates prompted by the hypotheses advanced by Lamarck and then Darwin). Thus ‘theorizing’ has been sometimes spelled out as ‘speculation’ which, in this context, has very often a pejorative sense (Callebaut 2013). Part of this view is certainly due to a general *envy towards physics*, seen as the paradigmatic example of how a science should be. Biologists then have tried to show that their discipline is as much ‘scientific’ as physics but nevertheless presents some irreducible features (Mayr 1982, 2004). In this sense it is important to recall the role of early philosophy of science (the Vienna Circle and Popper in particular) in shaping the image of physics itself and more in general of science (what is science and what it should be). Indeed, most of the debates among the nature of theories in philosophy of science stemmed out from reflections on physics and by people with, mainly, either physical or mathematical background. As recalled by Conrad Hal Waddington while “theoretical physics is a well recognized discipline […]” and “it is widely accepted that theories of the nature of the physical universe have profound

consequences for problems of general philosophy", on the contrary "theoretical biology can hardly be said to exists as an academic discipline" and "[t]here is even little agreement as to what topics it should deal with, or in what manner it should proceed" (Waddington 1968, p 525). No matter whether we adopt a syntactic or a semantic view, theories are thought to have a unification power. More, they are generally conceived as the main way to put order (an explanatory order) into the messiness of common experience. Thus it should not surprise that Waddington himself thought about theoretical biology not just as the systematisation of a particular biological process but rather as "an attempt to discover and formulate general concepts and logical relations characteristic of living as contrasted with inorganic systems" (*ibid*). The specificity of such a distinction is quite interesting as it is intertwined, in a different manner, with different forms of reductionism. On the one hand, Waddington seems to envisage the necessity of a theoretical apparatus specific to biology and its peculiar object of inquiry: life. On the other hand this move should bring biology closer to physics, at least in terms of its epistemic structure. Indeed the relation with physics is quite peculiar in itself. Surely physics played a decisive role in the rise and development of molecular biology (see for instance Morange 2000) by fostering the idea of the chance of 'explaining life' through its formidable theoretical apparatus. However the contact with biology challenged some theoretical physicists till positions like Bohr ones on the necessity of extending natural laws in order to give a scientific account of life (see for instance the volume edited by Sloan and Fogel 2011). In any case it is manifest that "theoretical biology is a highly heterogeneous type of enterprise, not only because it applies to widely divergent sub-fields of investigation – from genetics

and molecular biology to ecology and evolution – but because it proceeds via the application of a panoply of methods, which in some cases yield contrasting insights or highlight fundamental conceptual differences in the ways different theoretical biologists think of their subject matter" (Pigliucci 2013, p 291).

Biology and the problem of theoretical unification

Despite all these distinctions, many scientists themselves express the need for theoretical unification in the life sciences (see for instance Brenner 2010). The urge for theory is generally motivated by some considerations about theory's features. As already mentioned, theory should indeed have a unification power over a discipline, giving sense and order to the diversity of empirical findings. Moreover it should provide a general explanatory framework in which single results take place in a structured way (in this sense see also Kitcher 1989). However, on the other hand, some philosophers (see for instance Dupré 1983, 1993 and works of other scholars of the so called Stanford School) have argued against the unity of science, precisely claiming that such a unity is both *a posteriori* reconstruction and a misrepresentation of scientific practice itself. I suppose that, proverbially, the truth lies in the middle. Disunity of scientific practices is certainly evident in many areas of research. But it is also honest to recognise that unification attempts have been made throughout the entire history of scientific disciplines (again, in the life sciences, let us think about Modern Synthesis as a way to put together, consistently, Mendelian genetics and Darwinism). In addition, the very notion of *unity of science* can be spelled out in different ways (often interweaving ontological and epistemological levels). In ancient times, while Plato argued for a general theory of knowledge that should

have unified different branches, Aristotle claimed for a structural harmony of the different disciplines which should be connected to each other but cannot be reduced one to another. In modern age, Kant argued that the unity of different scientific practices would occur at the epistemic level: unity is *a priori* principle of reason guiding the process of scientific discovery by setting its conditions of possibility. Later on, logical empiricists put the question of unity of scientific knowledge on the top of their research agenda thus arguing for a methodological unification and vindicating the commitment to a united and consistent logical form. Since the aim of logical empiricists was profoundly foundational, they did not claim however that the content of different and specific theories and concepts should be unified but rather they thought about mathematical logic as the way to highlight the common formal structure in which all the sciences, hierarchically depicted, would be embedded. The debate is vast and other distinctions can be made. For instance, Popper has in mind the notion of unity as a tool to identify science from non-science, he argued in favour of unity of a unifying methodological stance: *falsificationism*. On the contrary Carnap was more concerned with a notion of unity that aims at building a general and comprehensive system of science. Recently Margaret Morrison (2007) has nicely argued that theoretical unification and explanation are not necessarily coupled together, thus meaning that a common, encompassing conceptual structure does not provide, *per se*, a general explanatory framework.

Therefore, the problem that scientists are facing now, is then how to define and then arguably to build, a unifying tool, given also the fragmented picture of biological research I have sketched so far. In this aspect, I think, lies the connection with the implementation of classificatory tools and computational

methods building general maps of biological data and biological knowledge. If some argued, especially within scientific journalism, that such practices would have decreed the *end of theory* (Anderson 2008), other scholars have claimed, on the contrary, that these approaches would have actually build, finally, a common theoretical ground for the life sciences (Ashburner *et al.* 2000). Of course these proponents had in mind, more or less explicitly, a different, compared to previous attempts, account of theory itself. Again, I would argue that it is the practice that shaped the way 'theory' has been addressed in this context. Speaking metaphorically, as in polytheism, there is not a single event generating all the *deities*. Each divinity has its own origin, story, myth. The many 'mothers' of contemporary biology urged to be treated equally and, in a sense, independently. After more or less three centuries since the release of the *Systema Naturae* by Linnaeus, the question is not just about building a new, updated version of the system but also whether the system itself is the right intellectual tool to forge such a unity. The ramification of biology itself and diversity of biological objects have probably a strong responsibility in the development of the taxonomic style. Such a *tropical rainforest* seems to resist to any form of overarching unification. Surely there is evolution as a general background stance. But, again, a precise theoretical unification is far from being achieved. Forms of life came to be so different that cannot be completely reduced one to another (see Carroll 2006). The inner diversity of the living things must be kept rather than neglected in any unification attempt. Thus, following Linnaeus' inspiration, if a unification is to be found for biology, it seems it must be from the bottom, from the collection of such diversities. On the other hand, contrary to Linnaeus, it is not clear whether such a bottom-up theoretical work would create a system.

Probably, classifying as a style of reasoning requires its own stance on what theory is. Actually, as Mueller-Wille (2007), Strasser and Chadarevian (2011) and Leonelli (2012) have shown, the classificatory practice in biology was and is not theoretically neutral. It is a matter of style. A question of how things are done and thought they should be done. Indeed "[t]he way scientists collect and order data is shaping their research since it is shaping the type of questions they pose and how they pose them" (Boem, Boniolo and Pavelka, 2015). Again, following Werner Callebaut suggestions, to understand what theory is in this context, we should not focus on a particular conceptual account, but rather concentrate on the practice of making theories by scientists (Callebaut 2013). In other words we should shift from 'theory' to 'theorizing'. Therefore the point is not to develop a particular account of theory and then trying to verify whether it would fit with the practice of research but rather to examine what counts as theoretical for practising scientists. According to such a perspective, it is central thus to evaluate the contemporary need for mapping as how it stimulated researches to look for links among different areas of research in a way that preserves a kind of uniformity. "Progress in the way that biologists describe and conceptualize the shared biological elements *has not kept pace with sequencing*. For the most part, the current systems of nomenclature for genes and their products remain divergent even when the experts appreciate the underlying similarities. Interoperability of genomic databases is limited by this lack of progress" (Ashburner *et al*. 2000, emphasis is mine). In other words it seems that due to its fragmentation, contemporary biology is facing a situation in which research is pursued by adopting many different languages but no translation devices.

Contemporary databases, are indeed maps of specific aspects of biological understanding but it is not clear whether they should be situated in relation one to another and thus in the general fabric of knowledge. If map thinking is employed to build biological maps dealing with specific empirical data, then maps of maps, *metamaps*, are needed to situate those data in relation to phenomena of interest. To put it differently, if this context knowledge can be often coupled with ordering. Given such circumstances, the adoption of a normalised, computationally controlled vocabulary, able to bridge the boundaries between distant areas of investigations, has been invoked as an effective way to *integrate* the results coming from disparate experimental settings, model organisms, technical devices, practical and theoretical aims. The problem of integration has not been addressed so much within the philosophical literature with few exceptions (see for instance O'Malley and Soyer, 2012). However a precise taxonomy of what is integration in computational biology is still unclear. Data-integration is generally considered as a technical problem due to the intrinsic differences among the procedures of data production and collection. However, given the heterogeneity of data, especially in biomedical research, not just one approach has been provided. Some scholars suggest that, generally speaking, integration "encompasses the combination of methods and methodologies [...], the process of making data-sets comparable and re-analysable, and the variety of ways in which explanations are brought together in a particular inquiry" (O'Malley and Soyer, 2012).

Data-integration should also be distinguished from methodological and explanatory integration. While methodological integration is generally applied to the context of systems biology as the way to obtain a "multidimensional

understanding" of a particular network or system by integrating different techniques, explanatory integration concerns the unification of different theoretical contributions in order to explain a phenomenon or a set of phenomena and the use of models in a specific field from another area of inquiry. On the other hand data-integration "refers to the process of theorizing and modelling databases, quantifying data accurately, developing standardization procedures, cleaning data, and providing efficient and user-friendly interfaces to enable data not only to be reused, but also reanalysed and combined in novel ways" (O'Malley and Soyer, 2012). On this aspect Sabina Leonelli (2013) has recently proposed a further distinction among different kinds of data integration. The first one is what she calls "inter-level integration" which concerns data representing different aspect of a single species in order to develop a multidisciplinary and whole-oriented understanding of such a species. The second one is named "cross-species" integration which compares data coming from different organisms in order to clarify the common biological ground of certain mechanisms shared by several living beings. In the end the third one, "translational integration", which refers to data of different sources (not only those coming from academia but also those ones from scientific institutions, companies etc. etc.) in order to provide "interventions to improve human health" (Leonelli, 2013).

In order to provide an epistemic analysis of these ways of integration and unification it is necessary to briefly reconstruct the origin of biological databases and their conceptual roots.

Unification in biology: from databases to bio-ontologies

The molecular revolution and the subsequent increase of information boosted the creation of repositories and databases. However, the creation of databases is independent from their adoption in biological research. Very often their origin is traced back to developments in computer science and the need of data management: namely how to label and display, in an ordered and easy to retrieve manner, the information of interest. As claimed in a quite recent handbook of computational biology "[b]*iological knowledge is stored in global databases*. The most important basis for applied bioinformatics is the collection of sequence data and its associated biological information" (Selzer, Marhöfer and Rohwer 2008, p 45, italics is mine). According to the type of data stored, it is possible to distinguish several kinds of biological databases. "Primary databases contain primary sequence information (nucleotide or protein) and accompanying annotation information regarding function, bibliographies, cross-reference to other databases" (*ibid*). Famous primary databases are GenBank (an American database for nucleotide sequences), EMBL and DDBJ (which, respectively, the European and the Japanese counterparts to GenBank), SwissProt and UniProt (both are repositories of annotated protein sequences, but the latter combines info coming from the former with other databases). On the contrary, secondary databases display a second-order information, summarising findings and analyses based on information kept in primary databases. Databases keeping information about literature (research articles and other publications) are also considered secondary databases. Examples of secondary databases are Interpro (that integrates different secondary information in a uniform system) or PDB (which is a repository of crystal structures of macromolecules). Biological databases are

often built and structured according to *relational models* which are in turn based on first order logic and notion of *relation*[13] in set theory. The basic assumption of relational models is that *data* can be represented as relations. In order to do so, specific *tables* are constructed and linked one to another via unique *key words* connecting data across different tables. For example, by taking the *fiscal code* of a person as a common key for different tables, it is possible to display information both on the physical domicile and on the health profile of that person.

Despite its technological novelties and contrary to naïve intuitions, the source of this practice is ancient. Behind modern database there is a precisely traceable philosophical vision. Collecting the world in order to understand it is an old, powerful idea. It is a style of reasoning. In the 17th Century the problem of classification was not just and simply to attribute names to plants and animals, it also involved the creation of specific instruments to pursue such achievements. As I said before, any classificatory effort is loaded with conceptual stances. Let us examine them more in details.

First, as Paolo Rossi recalls (1997), the issue of classification implies, explicitly or not, that a theory about the structure of nature is put in relation with a theory about the structure of language. This should not be intended naïvely as if knowledge simply stems out from naming. Rather, it represents the theoretical stance that a precise and consistent way of labelling things would be capable to grasp the actual classification in nature (to carve nature at its joints). Again this would not be achieved in virtue of the choice of this or that word, but according to the way and the coherence terms are in relations to each others. This aspect is

[13] Roughly speaking, a *relation* is a subset of the Cartesian product between two or more sets

somehow crucial as it sets up the basis for a philosophical perspective in which semantics and ontology are coupled together (see for instance Kripke 1980 and his notion of *rigid designator*). I will discuss the implications of such a stance further in this study. Second, by classifying scientists do not just deal with knowledge, they also cope with mnemonic concerns. Third, the language adopted by classifiers is able to grasp and elicit what is important over a myriad of details. All these dimensions are intrinsic features of the taxonomic style. All these points fostered the request and creation of artificial, universal languages, that should overcome the intrinsic limitations of natural one. The quest for a *Lingua Philosophica* can be seen as a very structured attempt to provide a unification tool for science. The fact that it is deeply related to the taxonomic style is not by accident. For instance, Leibniz's idea of a *Characteristica Univeralis* (a forerunner of modern logic) as formal and complete language to describe mathematical and scientific concepts goes precisely in this direction. By the second half of the 17th Century, scholars of various interests (from philosophers and mathematicians to linguists and taxonomists till polymaths as Leibniz himself) began to develop different proposals on artificial languages and formal tools precisely to cope with the issue of the classification of nature. The formal aspect here does not always mean that all these approaches were formalised. Contemporary formal logic is not as such because it has symbolic formulas, but rather because it highlights the *form*, the structure of a sentence. It grasps its 'essence'. Thus even if, for instance, Linnaeus adopted Latin, it is the way the terms were structured, their formal relation, that granted the effectiveness of its classificatory power.

All these artificial languages, in spite of differences, show some common important features (see Rossi 1997). First, the terms do not correspond directly to things in the world, but to the concepts that refer to those things. Second, the connectives and symbols of formal languages should highlight relations and connections among the objects represented. Third, exactness is demanded so that semantic ambiguity is avoided. Fourth, such a universal language should imply the realisation of a universal knowledge, meaning the complete enumeration and classification of all those entities that have a term referring to them. This means also that this completeness must be intended in a strong epistemic sense, given that the limits of such encyclopaedic enterprise are indeed the limits of the language itself. According to Rossi (1997) this is also the perspective into which one should interpret Francis Bacon's *tabulae.* Rossi thinks about Bacon proposal as an anticipation of a sort of database, meaning that those *tables* (think indeed about the aforementioned relational models) were actually built as a way to display knowledge in an organised way capable to let differences and similarities emerge. In this view knowing is precisely naming and ordering correctly. This approach is not just a philosophical stance, conceived in principles, but it also mirrors itself in the practice. As also claimed by Linnaeus himself: "[f]*undamentum botanices duplex est: dispositio et denominatio*" (Linnaeus, 1751, 151). Moreover, previous botanical classifications during the Renaissance, based on the work of Galen or Dioscorides, usually ordered plants according to their medical and pharmacological relevance, thus putting botany as ancillary to medicine and, more importantly, bending classification itself to a particular interest rather than trying to find the relations among things. New classificatory strategies and languages were thought to be able precisely to get the essential

over the accidental. Here lies the importance of their capacity to reveal the formal structure. At the end of the 17th Century, the great French botanist Joseph de Tournefort used 15 words to describe the bulbous bluegrass which is, in Linnaeus system, *Poa bulbosa*. As Rossi (1997) argues, in those 15 words there is less information than that in the two ones adopted by Linnaeus.

It is not a mystery then that the discipline concerned with information, *informatics*, stemmed out of logic. Alan Turing and Von Neumann thought about *computers* precisely as mechanical and automatic replicas of individuals, capable of tremendous computational abilities. Thus it should not surprise that computer science dealt with data classification and management since its beginning. Indeed, ordering and classifying the world is then not just a mode of representation but also a heuristic procedure. By collecting and comparing it is possible to produce and foster discovery. The implementation of computational instruments in biology, as I tried to show in the previous chapter, did start after the second half of the 20th Century, many years before the rise of computational biology. The development of the first modern biological database, the *Atlas of Protein Sequences and Structure* by Margaret Dayhoff in the 60s (see also Strasser 2010) and the discovery of sequencing by Frederick Sanger in the 70s can be considered two iconic moments, standing as the first steps to such a change of research practices. In the late 80s, once distinct databases started to be constructed and implemented (*e.g.* GenBank in 1982, SwissProt in 1986. *Nota bene*, in the same year the term *genomics* has been coined) a need for a common groundwork became urgent and precisely from a practical point of view. The increase of information, due to mapping efforts of molecular findings, required to be addressed. "Keeping pace with molecular developments were biological data-

management efforts" (Lewis 2005). In the end, the complete sequencing of several model organisms at the end of 90s furnished the biological basis of a common ground for building bridges among research communities. These diverse biological maps could finally be compared. In particular, it was necessary to find a practical solution to treat gene functions in a consistent manner that nevertheless would satisfy the specificities of the diverse organism models. Both from social and epistemic point of view, the choice of the experimental model shaped the diverse research communities. These communities were not based only on organism-specific researchers, but they also incorporate scientists with different backgrounds as biochemists or geneticists. Nevertheless the model adopted also determined to which community a researcher belonged, framing and tailoring his/her research language. This is because "many biological databases bloomed, flourished and […] all of them operated primarily autonomously" (Lewis 2005, p 103). In order to cope with such a *Babel Tower*, the development of a shared language was perceived as fundamental. Similarly to what happened in the 17th and 18th Centuries, one solution proposed was the adoption of a formal tool, that should highlight formal structures and relations. Thus, in those years, some biologists started to look at the formal work of data managers and computer scientists precisely as naturalists of the past, embarked philosophers, linguists and logicians in helping them solving their practical and theoretical issues. In 1998, Michael Ashburner (professor in Cambridge and developer of FlyBase, a repository of information concerning *Drosophila*) proposed a simple, hierarchical, controlled vocabulary for representing gene function among different communities. Following his vision, representatives of different model organisms databases, (initially FlyBase, Saccharomyces Genome Database, and

Mouse Genome Informatics) met together and agreed to a common framework for labelling and characterising the functions of genes. This collaboration constituted the first basis for the creation of the Gene Ontology (GO) Consortium.

Before exploring GO more in details, analysing its strengths, limitations, scopes and structure, it is necessary to examine the relation between computational ontologies and ontology in a philosophical sense. Despite the fact that several bio-ontologies are built in the total unawareness of what ontology is in philosophy, there are both historical and theoretical reasons to support the argument that actually ontology played, and still plays, a critical role in shaping the work of engineers in knowledge representation. As Daniel Dennett wrote once "[t]here is no such thing as philosophy-free science; there is only science whose philosophical baggage is taken on board without examination" (Dennett 1995).

CHAPTER III

Ontology and ontologies

In philosophy the term 'ontology' is usually adopted to design that area of theoretical speculation involved in the analysis of what there is and the nature of being. Nowadays, sometimes the term is equated to *metaphysics*, while Aristotle called it *first philosophy* as he meant that such a reflection on the more abstract categories and relations should come first and prior to any other knowledge (since it would apply to any science). Edmund Husserl (1900/1) provides a further distinction between *formal* and *material* ontology. While the former deals with being *qua* being, the latter is linked to specific areas/regions of reality or is concerned with their representations in given theories. Quine (1953) thought that the only, genuine, approach to the ontological problem would be through the analysis of the ontological commitment of scientific theories. In other words, according to Quine "the ontologist's task is to establish what kind of entities scientists are committed to in their theorizing" (Smith 2003 p 3). Quine's turn had a strong influence in the way philosophical ontology has been conducted and defined. Since then, many philosophers (especially within the analytic tradition) stopped from seeking *a priori* true principles of reality, and they rather moved to look for those assumptions considered as valid according to certain theories or field of inquiry. After Quine, it seemed that the formal part of ontology gave the way to logic and epistemology and, especially, to science itself. This is also because, although often implicit, the determination of the fundamental entities and processes of reality rests now on natural sciences. The Quinean approach then somehow 'dissolved' the classical ontological problem by claiming that the

content and the structure of reality would have been revealed by empirical sciences while its representation would have been pursued by first order logic.

However, a return of the pretension of ontology in a more traditional sense started with Saul Kripke (1980) who built (or better, restated in a new shape) a tight connection between ontology and semantics. Thus he originally argued in favour of a link between the structure of the world and the structure of the tool we adopt to predicate about the world: the language. Famously, and contrary to classical descriptive theories, Kripke proposed a causal theory of reference, according to which a name stands for a thing in virtue of a causal connection with the thing through a procedure of *baptism* within the community of speakers. Names are then *rigid designators*. Moreover, identities between terms discovered by empirical sciences, such as *water* and H_2O, constitute then *a posteriori* necessary truths. Contra Quine, Kripke has shown how ontology can be put again prior to natural sciences. Again, despite Quine efforts against Plato's and Aristotle's beards, a rigid designator is a modern way to reaffirm the old notion of *essence*. Indeed, Kripke's legacy favoured the flourish of metaphysics and promoted the idea of a genuine ontological research independent from epistemology (and maybe prior to it).

At first glance, all this debate may seem extremely abstract, detached from scientific practice, and concerning precisely that type of speculation that irritates biologists so much. In a sense, it could be rightfully argued that applied ontology seems to follow Quine's ideas, whereas the disclosure of the ontological structure is a task just for natural sciences. However I would argue that the picture is more complex and complicated. Indeed, the very choice of the term 'ontology' to designate knowledge representation is not accidental and certainly,

the fact that semantic tools were considered adequate to elicit the real structure of world, slants on the Kripkean side. Even if engineers and computer scientists were not fully aware of the philosophical debate, they somehow bumped into it. They were trying to solve a practical problem. However such an issue dealt with the relation between the world and the language. Their question was indeed the *problem of reference*. As a matter of fact, in information and computer science the "task for the new 'ontology' derives from what we might call the Tower of Babel problem. Different groups of data- and knowledge-base system designers have their own idiosyncratic terms and concepts by means of which they build frameworks for information representation. […] Methods must be found to resolve the terminological and conceptual incompatibilities which then inevitably arise" (Smith 2003 p 6). Thus the very problem is not in naming (which is often arbitrary) but in the capacity of a structured nomenclature to grasp fundamental relations among things. This is not just a philosophical, speculative question. It pertains to real scientific research. In a tit for tat appeared in *Genome Biology*, first Sydney Brenner (2002) engaged the whole project about the creation of a computationally controlled vocabulary, claiming that it was a waste of time since terms are just words and not things (and things are what a biologist should care about). Then Lawrence Hunter (2002) replied by arguing that Brenner failed to understand what is at stake. According to Hunter, Brenner misses completely the point since the power of bio-ontologies "is not in the list of names they embody, but in the relationships they represent" (Hunter, 2002 p 2). Moreover, Hunter argues that such a rigid tool is not made for scientists but for machines. Ontologies allow computer programs "to accomplish complex inference tasks" (*ibid*). As in a modern Porphyrian tree, computational ontologies are displaying a

representation of the world. Once a term had been chosen and fixed, what it would stand for? To put it differently, what is at stake here, is the nature of the categories implemented in this representational work. Surprisingly, at least for modern scientists, the question is a sort of a new *dispute on universals*. And also in this contemporary, scientific debate, it is still possible to envisage nominalists versus realists. However, the purpose of applied ontology is not to solve a philosophical problem. Nor the aim of computational ontologies, although they are created to grasp something which is real, is to *carve nature at its joints*. This is because, paradoxically, the *nature* represented *at its joints* must not be intended ontologically. Surely many scientists are realist in this sense. Most of them believe in the categories and in the objects of their theories. They have pragmatic reasons to do so. Nevertheless ontologies pertain to the epistemic side. This realism is then always within a specific theoretical setting. Such a thing is not just for philosophers. As Gruber clearly states "for AI systems what 'exists' is that which can be represented" (1993). Thus it is the terminological choice, the type of design and format adopted, that shape and determine not only which entities exist but also establish that such entities have only those properties properly represented. Ontological work is close in this sense to modelling. This modelling however, it is not about this or that specific entity, but rather on what is an entity, and what should count for it in our domain of interest. As for philosophical ontology, computational ontologies come first indeed. They set up the objects and rules to any further analysis or research. As both a dictionary and grammar book, ontologies establish which are the 'words' that can be used in research, their meaning and the syntax. This is because for the ontological work, the domain of a scientific inquiry is not simply a given. As a matter of fact,

experience comes all together. However, knowledge requires distinctions. It is necessary to abstract some parts from others. An ontology, in a computational sense, is a way to do so. And there are many kinds of ontologies because "reality is like cheese: it can be cut in many ways" (Grenon, Smith and Goldberg 2004).

Upper-ontologies and bio-ontologies

In the field of applied ontology it is common to distinguish between *upper-level ontologies* and *domain ontologies*. Drawing an analogy with the philosophical distinction formulated by Husserl (1900/1) between formal and material ontology, upper-level ontologies deal with abstract and general notions such as 'object', 'process', 'relation', 'part' and 'whole', while the second ones implement specifications relative to particular domains of interest such as (in the life sciences) molecular biology, anatomy, diseases etc.

Bio-ontologies then, are a clear example of domain ontologies. This means that the fundamental categories deployed by bio-ontologies as substrate of their categories are not directly specified. Although not so much considered by both wet and computational biologists, upper-level ontologies are critical to understand the structure of bio-ontologies and their rationale. The Basic Formal Ontology (BFO) developed by Barry Smith is a good and famous example of an upper-level ontology. An ontology as BFO is indeed the type of tool to instruct the main elements on which one can develop different domain ontologies. First, BFO distinguishes all entities into *continuants* (objects) and *occurrents* (processes). "Intuitively, the big divide in the BFO lies between entities in three-dimensional space (continuants) and entities in four-dimensional space, *i.e.*, in space and time (occurrents). In biomedicine, this is like the difference between a

three-dimensional anatomical object such as the heart (a continuant), and the physiological functioning of the heart to pump blood (an occurrent). In a sense, continuants and occurrents represent two different ways of viewing the same objects (Robinson and Bauer 2011). To put it differently, while a *continuant* is conceivable as a snapshot of the world (*e.g*., an image of the liver), an *occurrent* is more like a movie view of reality (*e.g*., a video of the hepatic activity).

Philosophically speaking, BFO is definitely based on a sort of Aristotelian framework. Thus it distinguishes between general classes or types, called precisely *universals*, and particular instances of those classes, named *particulars*. The Aristotelian stance is also embedded in the peculiar understanding of scientific research adopted by BFO. Following Aristotle's argument that "there is no science of the individual as such" (*Met*. XIII, 10, 1086 b, 33) and nevertheless "our knowledge of the individual precedes our knowledge of the universal" (*Nic. Eth.* VI, ii, 1143 b, 5), such ontologies are constructed by adopting the view that scientific efforts deal with individual phenomena in order to construct hypotheses, claims, and theories about universal classes. In other words, let us consider a lab investigating how, a protein or a family of proteins of interest (as the Cyclin family proteins) play a role in a biological process (*e.g.* cell cycle) in a particular organism (*e.g. Saccharomyces cerevisiae*). Surely the lab would not examine all the yeast cells in the world but only a certain number of them (*i.e*. instances or particulars). Nevertheless, the conclusion of that research is that Cyclins are involved in cell cycle in yeast. Neither just this particular cell nor those cells are considered. Scientific ontologies represent universals. The philosopher of science understands pretty well how this view does not require a strong ontological commitment. This is a way of representing knowledge. A way

that must be effective and useful. Again, computational ontologies are not a proposed solution to metaphysical problems.

Finally, these categories must be connected. BFO display different kinds of *formal relations*, affecting both particulars and universals. Terms in computational ontologies normally refer to universals but in some situations certain relations regard particulars. The *is_a* relation denotes a relation between universals. For example, "the cofactor transporter activity *is_a* transporter activity" (Robinson and Bauer 2011, p 144). Instead the relation *instance_of* denotes a relation between a particular and a universal. For instance this yeast cell *instance_of* yeast. Finally the part_of relation stands for a relation between two particulars as this nucleus *part_of* this cell.

BFO is not necessary the unique ground on which one could built domain ontologies as bio-ontologies. However it provides a very well example of how such ground could be constructed. As in any encyclopaedic and classificatory effort, the construction of a common, shared, language is considered the first condition to be satisfied in order to achieve any progress. Again, history of ideas taught us how any effort of building a complete encyclopaedia is doomed to failure for its very constitution. However, as already said, contrary to philosophical ideals and intellectual agendas such as the *Characteristica Universalis* of Leibniz or the more recent research enterprise concerning the foundation of mathematics instructed by Frege and Russell, the theoretical horizon of computational ontologies is to fulfil a very specific pragmatic purpose.

Accordingly, especially within the life sciences, in recent years many efforts have been made to build a common platform of such a kind. The Open

Biomedical Ontology consortium (OBO) has been created precisely to implement a common ground for the proliferation of different biomedical ontologies. This necessity is critical not only as the sources of biological information (databases) are different, differently constructed and displayed, but also because their ontological representation is also diverse, concerning distinct research needs and levels of granularity. "In 2001, Ashburner and Lewis initiated a strategy to address this object level question by creating OBO, an umbrella body for the developers of life-science ontologies. OBO applies the key principles underlying the success of the GO, namely, that ontologies be open, orthogonal, instantiated in a well-specified syntax and designed to share a common space of identifiers" (Smith *et al.* 2007). OBO consists of many different ontologies, and is both technically and financially sponsored by the NIH Roadmap National Centre for Biomedical Ontology (NCBO). OBO provides also a common formal structure. *OBO file format* is "an ontology representation language. The concepts it models represent a subset of the concepts in the OWL description logic language, with several extensions for meta-data modelling and the modelling of concepts that are not supported in DL languages" (http://oboformat.googlecode.com/svn/trunk/doc/GO.format.obo-1_2.html). Moreover, the rapid increase of "ontological work" in the life sciences does not exhaust itself in the production of such a common ground. More actively, besides OBO platform, some ontology developers have started to assemble OBO Foundry, "a collaborative experiment based on the voluntary acceptance by its participants of an evolving set of principles (available at http://obofoundry.org) that extend those of the original OBO by requiring in addition that ontologies (i) be developed in a collaborative effort, (ii) use

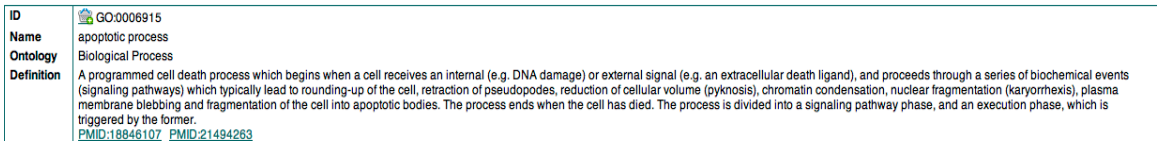
common relations that are unambiguously defined, (iii) provide procedures for user feedback and for identifying successive versions and (iv) have a clearly bounded subject-matter (Smith *et al.* 2007). OBO Foundry displays ontologies dedicated to different aspects of biological research, also cutting reality at diverse levels. Beside GO (on which I will spend more time later in this chapter), there is OBI (Ontology for Biological Investigation) which models the experimental design, research protocols, materials and methodologies, data used and the kind of analyses operated, CHEBI (Chemical Entities of Biological Interest) which is a standardised and unified classification of chemical substance of biological relevance or PATO (Phenotypic Quality Ontology), which is devoted to connect other ontologies (e.g. those coming from GO) to phenotypes represented as qualities/properties.

GO: an orienteering tool for biomedical research

Gene Ontology is probably the most famous ontological initiative developed for biological research. Gene Ontology aims to provide a standardized representation of gene products' features across different species and databases. GO actually covers three domain ontologies which are called *Cellular Component* (the parts of a cell of its extracellular environment), *Molecular Function* (the basic activities of a gene product) and *Biological Process* (the set of molecular events characterized by clear beginning and end).

GO *terms* describe gene product characteristics in a single, computationally controlled way, in order to provide a common format. Each GO term (Fig.1) has a specific name which designates it and which can be a single word or an expression (*e.g.* apoptotic process), a unique alphanumeric identifier (*e.g.*

GO:0006915), a definition (see the note[14]) with references, and the ontological dependence that indicates the domain to which it belongs to (*e.g.* Biological Process).



ID	GO:0006915
Name	apoptotic process
Ontology	Biological Process
Definition	A programmed cell death process which begins when a cell receives an internal (e.g. DNA damage) or external signal (e.g. an extracellular death ligand), and proceeds through a series of biochemical events (signaling pathways) which typically lead to rounding-up of the cell, retraction of pseudopodes, reduction of cellular volume (pyknosis), chromatin condensation, nuclear fragmentation (karyorrhexis), plasma membrane blebbing and fragmentation of the cell into apoptotic bodies. The process ends when the cell has died. The process is divided into a signaling pathway phase, and an execution phase, which is triggered by the former. PMID:18846107 PMID:21494263

Fig.1 (taken by QuickGO)

Each GO term has then a set of defined relationships (*e.g. is_a*, *part_of*, or *positvely_regulates* etc.) towards one or more terms in the same domain, and sometimes in other domains. The GO terminology is designed to be species-neutral, in order to be exploitable from prokaryotes to eukaryotes and from single to multi-cellular organisms.

GO annotation is "the practice of capturing the activities and localization of a gene product with GO terms and it provides references and indicates what kind of evidence is available to support it" (GO website - http://www.geneontology.org/). Annotations are created on the basis of observations of the individual occurrences (*i.e.* the instances) of the type under examination. Hidden in this scientific and technical presentation, the philosopher may recognise the Aristotelian mark of such an endeavour. Indeed, while GO

[14] "A programmed cell death process which begins when a cell receives an internal (*e.g.* DNA damage) or external signal (*e.g.* an extracellular death ligand), and proceeds through a series of biochemical events (signaling pathways) which typically lead to rounding-up of the cell, retraction of pseudopodes, reduction of cellular volume (pyknosis), chromatin condensation, nuclear fragmentation (karyorrhexis), plasma membrane blebbing and fragmentation of the cell into apoptotic bodies. The process ends when the cell has died. The process is divided into a signalling pathway phase, and an execution phase, which is triggered by the former"

terms stand for types, GO annotations are singular evidences (obtained through experimental observations) that instantiate the term of relevance. Here lies the Aristotelian legacy. Knowledge, biological knowledge, belongs to *universals*. However it is possible to get to the universal through the *particular*. GO annotations display the gene product (*e.g.* PB1-F2 protein), the relevant GO terms involved (*e.g.* apoptotic process), the reference which provides ground for such an annotation (*e.g.* the Gene Ontology Database references), the type of scientific evidence that supports the annotation (*e.g.* Inferred from Electronic Annotation) and finally the author and the date of the annotation itself.

It is clear that the choice of the three domains is also motivated by reasons of convenience. In other words, since GO is meant to provide a semantic representation of knowledge in use for molecular biology, the conceptual framework adopted clearly refers to the way molecular biologists pursue their experimental work, display their information and conceive explanations. This illustrates why GO is built to present *terms* and *annotations* according to a mechanistic description of molecular events. Indeed GO is a technical tool, not a metaphysical device. Its application reveals the reason behind the terminological choice. However such a choice, given the scope and the hope for generality of GO, cannot be grounded just on logical consistency and empirical adequacy. Being a tool of knowledge-capture and representation, GO terms must satisfy the needs and the desiderata of the scientific community. Accordingly, the process of *curation* is the production of annotations on the basis of findings retrieved from experimental work. Thus, since the activity of curation requires a deep scrutiny of the relevant literature, it is important (no less than obvious) that curators

possess a robust expertise in the related field. Normally, annotations are created through a procedure that requires several steps.

The primary aim of GO annotation is to create annotations based on findings obtained from experiments on related organisms. However information coming from different model organisms or by sources other than experiments (as sequence information in the genome browser) is also taken into account. The annotation file provides thus a way to discriminate the sources of annotation and to filter out what is not considered important by the researcher. As in a map, the single scientist can highlight this or that feature, remove or add elements, in order to orientate himself/herself in the topic.

The second step consists in linking the information captured by the annotation within the appropriate term. Some factors should be taken into account. Indeed the kind of experiment itself shapes the nature of evidence that can be obtained and sets up the resolution and the quality of results. "For example, cell fractionation might localize molecules of a protein to the nucleus of a cell, but immunolocalization experiments might localize molecules of the same type of protein to the nucleolus of a cell. *As a result, the same gene may have annotations to different terms in the same ontology because annotations are based on different experiments*" (Hill *et al*. 2008, emphasis is mine). Last, but not least, annotation procedures are usually verified for their consistency. In doing so, both computational/logical tools and domain experts are involved. To further develop this aspect, it is possible to individuate distinct *epistemic moments* according to which annotations are created. First, information coming from scientific publications is captured, extracted and abstracted by annotators and then condensed into a unique *semantic designation*, according to the rules of

term composition and the consistency of GO. Thus, even if most annotations are manually operated, the process is reviewed both by GO curators and by automatic reasoners. Such a product must finally face the judgment of the scientific community *i.e.* the experts of the field. Obviously, the process of annotation is not a static given. Both GO terms and annotations are in constant evolution and growth since they map the current state of the research. GO updates its content according to scientific debates and it is even able to display the disagreement among experts (*e.g.* the NOT annotation). For example, the vast part of terms and annotations pertaining to the range of phenomena which include the death of a cell are undergoing a revision due to the very last scientific finding in the field (see for instance Kaczmarek, Vandenabeele, Krysko 2013; Christofferson and Yuan 2010).

Gene Ontology then, is not dictating, in a purely top down fashion, which terms are right or not for the research, but it is rather mapping the current use of *scientific vocabulary* trying to standardize it. However such a feature shows why GO is also normative too. Indeed the standardization created through GO affects the way information in databases and other electronic resources is presented. By expanding the experimental context, ontologies allow not just the use but especially the re-use of the represented knowledge. Thus a new lab, in the definition of its standards and terminology, would not start from scratch, following arbitrary criteria, but it would rather rely on a body of knowledge which is the more and more organised and unified (I will come back on this point later in the study).

The structure of GO terms and relations among them is also displayed graphically (see an example in Fig. 2)

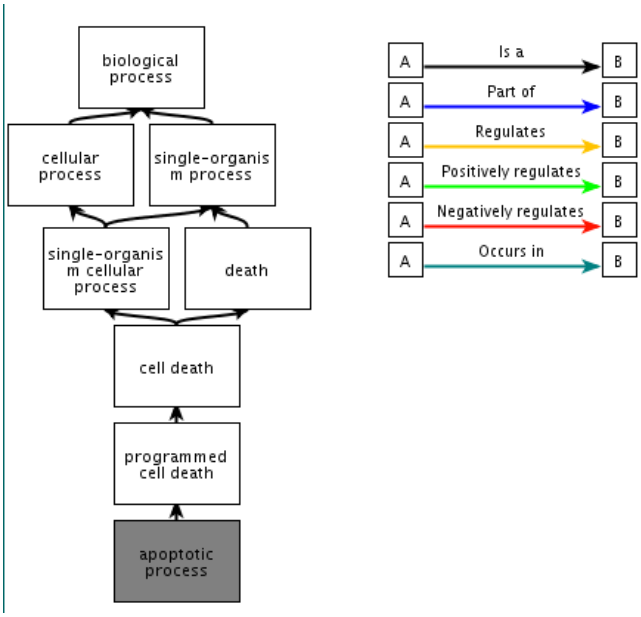


Fig.2 (taken by QuickGO)

This shows very well how GO is an epistemic map. A map of knowledge. Indeed each chart is highly interactive. Each term can be *opened* and further examined. *Ancestors* and *children* terms are thus shown along with related gene product annotations. All this information is literally mapped into a wider context. Therefore it is possible to navigate GO through its terms and relations, check the gene products involved in certain phenomena and link them with other area of research out of the given experimental context. Moreover GO is a live map. As already mentioned, the content of GO is not static but rather it tracks the changes and developments within the scientific community. Let us briefly consider a practical case. For a long time the term *apoptosis* has been considered a synonym of programmed cell death, opposed to *necrosis* which, in turn, referred to the death of a cell of accidental reasons. The idea of programmed cell death has been formulated in the 60s and then further developed in the 70s and 80s when it

has been labelled as apoptosis, suggesting the presence of a regulated cell death process common to all cells usually dependent on a family of proteins known as *caspases* (Vandenabeele, Galluzzi, Vanden Berghe and Kroemer 2010) opposed to forms of cell death considered purely passive and contingent. Since the late 80s new evidences indicated that also some necrotic phenomena should have been seen as regulated. In 2005, these new observations eventually led to the 'baptism'[15] of a new term, *necroptosis* to designate a form of regulated necrotic cell death (*ibid*). Underlying molecular mechanisms of necroptosis are still under investigation but the more and more evidence suggest they are radically different to apoptotic ones. The consequences of these discoveries for and on GO are crucial. First, the terms *programmed cell death* and *apoptosis* can no longer be considered synonyms. According to the state of the art, the apoptotic process is a kind of programmed cell death. Thus the entire architecture (of the whole cell death section of GO) and relations among these terms and the related ones should be revised. Indeed, 120 terms circa were introduced, 18 terms became obsolete and 200 terms were modified to better represent current scientific knowledge. To give an idea, this meant an overall work (mostly manually conducted) on more than 7000 annotations. As a result, the term *necroptosis*, that in 2009 counted only 10 related annotations, scaled up to 298 annotations in 2014 (personal communication during my vising period at the EBI).

Of course, such a revision process was not just a technical and automatic thing. GO curators had to update the status and the advancement of a particular biological knowledge by considering its impact on other areas of biological research. This means that the entire process involved several meetings, called

[15] just remember the Kripkean stance, see page 73

content meeting, between GO curators and leading experts in the field (in this case cell death), comparing their ideas and, accordingly, either promoting/proposing or hindering specific semantic solutions.

This aspect shows how much GO depends on epistemic interactions between different 'players'. Curators, annotators, computer scientists, researchers, field specialists (sometimes even philosophers) contribute all together, bringing their different point of view and expertise, to the maintenance and the development of the tool. Such a peculiarity is particularly evident when new terms are proposed or a revision of existing terms is required. During my visiting period at the EBI a researcher in the field of *Dinoflagellata* (a *phylum* of flagellate protists) asked for the introduction of the term *sulcus* that should stand for a particular *cellular component*: a cell surface furrow usually hosting one of the two flagella. Such a request promoted a debate within GO curators since the term *sulcus* already existed in GO for a completely different context[16]. Here the appeal to field experts can help only partially. As a matter of fact, despite the problem of consistency (the same term cannot mean two different things), GO curators have also been involved in discussions that, from an external point of view, could have seemed almost metaphysical. Indeed they asked to themselves, what a furrow precisely is and whether it could be defined as a type of hole or rather as a folded surface etc. Although in its purpose this work is purely descriptive, it should not surprise that some scholars (as Barry Smith) see this type of effort as a genuine ontological work as the GO developers would reveal true natural kinds. As already argued, I do not think that this is the case. Again, GO maps biological

[16] "The series of molecular signals generated as a consequence of a fibroblast growth factor-type receptor binding to one of its physiological ligands resulting in the formation of the lung bud along the lateral-esophageal sulcus" (QuickGO website)

knowledge and not supposed metaphysical categories about nature. In addition, GO does not map biological knowledge as such, but rather according to a formal structure that is certainly pragmatically convenient but there is not evidence that it refers to the ultimate essence of reality. Moreover GO three main ontologies are thought and constructed according to the most adopted type of explanation in molecular biology: mechanistic explanation. Because of this it should be clear that GO is definitely an epistemic tool.

Of course, despite Smith's derogatory attitude about epistemology, being epistemic does not mean less objective neither less important for science (even in its practice). On the contrary, the clarification on what epistemic ground GO provides its analysis could contribute, not just to the understanding of GO, but also to a better use of it. Indeed, once the nature of GO terms and annotations is more clearly specified, scientists would consider GO and its potential in a more proficuous way, and they would avoid behaviours that could led them to several biases. First of all, since GO is a map of what is already known, it is not, *per se*, a discovery tool. However GO creates the conditions to make discoveries.

Let us consider one of the most common operation available with GO: the *enrichment analysis*. This type of investigation may allow scientists to map and evaluate possible scenarios given specific experimental conditions. For instance for "a set of genes that are up-regulated under certain conditions, an enrichment analysis will find which GO terms are over-represented (or under-represented) using annotations for that gene set" (http://geneontology.org). By doing this, researchers can characterise that set of genes under a common functional profile, revealing important features of the underlying biological phenomenon. The output of such an analysis is then an ordered list of GO terms, with the related p-

value. For example, due to high-throughput analysis, it is now possible to compare the gene expression profiles of an healthy tissue with the cancerous one. By examining the semantic discrepancies resulting from the analysis it is possible to furnish indications about the differences on the hidden biological mechanisms. This kind of work can also be done pursuing different research strategies. On the one hand it is possible to check which terms are significant in a particular set of genes or the other way round, that is to check if a biological phenomenon (such as apoptosis) is over-represented (or under-represented) in a particular set of genes. Several tools (such as DAVID, Panther, Ontologizer, Onto-Express) have been developed to perform this type of investigation deploying different statistical methods and different databases sources. This means that researchers should ideally perform different functional profiling adopting different tools before interpreting their experimental results. Thus, GO does not provide any, *stricto sensu*, discovery since a map cannot show what is not mapped. However a map can let the interpreter to observe connections that are invisible without the map itself. This is exactly what GO can do.

Another kind of common analysis with GO consists in the prediction of putative gene function. "Typical approaches tend to be variations of the same theme: genes are grouped together on the basis of some criteria such as similar gene expression or through a protein–protein interaction network. Enrichment of GO terms is detected by methods such as those described above, and the uncharacterized genes are presumed to be involved in the same biological processes as the genes with which they are grouped" (Rhee, Wood, Dolinski and Draghici 2008). It is clear that such an operation pays close attention. By propagating a gene function just on the basis of annotations that are neither

manually verified nor experimentally validated can lead to many false positives. On the other side "[g]ene functions can also be inferred from GO annotations without the need for a prior gene grouping, for instance, on the basis of a semantic analysis of the gene function association matrix. This type of analysis relies on capturing the implicit dependencies that might be present between genes" (*ibid*). On this aspect some scholars (Noble 2008) have raised some critical points about the power and the limits of GO, by claiming that describing biological phenomena just in terms of their related, involved, gene products will miss the higher-level insight. In order to answer to this issue, many new projects in this field are devoted to the development of other ontologies, integrated with GO, but concerning different aspects and levels/granularities of biological phenomena.

To sum up, although GO is an extremely powerful tool, scientists who use it should be aware its features, paying attention to the nature of annotations and to the distinction between annotations and terms. All I have described, I believe that points very much to the fact that GO is concerned with the epistemic side. Again, by this, one should not deduce that GO does not speak about the real world. Moreover I think that scientists are not very much interested in that philosophical position known as *scientific realism*. Pragmatically, they usually behave towards the entities they deal with, by following Hacking's claim (although probably unaware of it) that "if you can spray them, then they are real" (Hacking 1983, p 23). In this sense it is also possible to appreciate how and why a tool like GO is *objective*. Peter Galison and Lorraine Daston have nicely shown that objectivity itself has a history (Daston and Galison 2007). Contrary to naïve intuitions, different periods of scientific development involved distinct notions of

objectivity. One form, particularly relevant for maps, is the so called *truth to nature* meaning that some images and representations were thought as the "closest rendering of what truly is" (Daston and Galison 1992, p 84). Indeed maps, atlases and databases fostered a particular idea of objectivity as they are able to perform a *standardisation* of the objects observed (through the common, relational, representation rules) and the subjects observing (via the common interpretation guidelines). Standardising means in fact selecting the working objects of scientific inquiry opposed to the unbearable variety of natural ones. For example, the anatomic atlases of the 18th and 19th Century are highly based on standardisation. Moreover such maps display their objects as *typological*. Indeed, these atlases show, for instance, "this typical liver rather than one with hepatitis" (Daston and Galison 1992, p 85). Standardisation and mapping activity must be taken together. As recently shown by Müller-Wille (*forthcoming*), when the great Danish botanist Wilhelm Johannsen started to use statistical methods in order to understand the mechanism of inheritance by *mapping* different phenotypes, he struggled a lot with the intrinsic variety of plants that rendered those samples unable to be properly treated from a statistical point of view. One of the first steps Johannsen adopted was precisely the 'purification' of the samples by creating 'pure lines', or varieties showing *stereotypical* features. Indeed maps allow comparison of data by displaying *typical phenomena*. Ontologists would say that maps deal with *universals.*

GO. From description to normativity

However there is definitely more. A tool like GO is a map of knowledge but it is also a way to standardise practices. Accordingly, GO presents a form of

objectivity that, following Alberto Cambrosio's suggestions (Cambrosio, Keating, Schlich and Weisz 2006), can be called *regulatory objectivity*. This kind of objectivity "is based on the systematic recourse to the collective production of evidence. Unlike forms of objectivity that emerged in earlier eras, regulatory objectivity consistently results in the production of conventions, […] most often arrived at through concerted programs of actions" (Cambrosio, Keating, Schlich and Weisz 2006, p 189). As I have previously shown, both GO structure and its practical choices, heavily rely on collective concerted actions among different 'players' such as database curators, biologists, other researchers, computer scientists etc. A map at first glance, is a standardised representation of different elements under a shared, common framework. Standardisation implies also agreed conventions both in the construction and in the interpretation of what is represented. In this sense a simple description might become a norm.

It might be useful to use an metaphorical image to explain this epistemic passage from descriptive efforts to normative ones. In the field of western jurisprudence, it is possible to individuate two main legal systems that have been developed differently. These two systems, known as *civil law* and *common law* are primary distinguishable because of their different historical genesis, thus affecting the countries in which they are applied, and then because they embed different conceptions concerning the nature of jurisprudence itself and thus the nature of what a norm is. Civil law, preponderant in all European countries (with the exception of the UK) is rationalised in the framework of the ancient Roman law system, further developed by the code of Justinian and finally systematised by the Napoleonic code and the German BGB (Bürgerliches Gesetzbuch). Accordingly, civil law is based on the written codification of general norms and

principles that constitute the primary source of law. Such systematic collections of principle, inform both citizens about the behaviour they should have and judges/magistrates on how interpret the law itself. Therefore, civil law establishes and explain, from above, principles, rights and duties and how the legal system works. Civil law founds and unifies jurisprudence by acting as a sort of a top down theoretical framework, thus determining what is consistent with its principles and norms, and rejecting what is contrary to it.

Common law is instead the system adopted by the UK and by most of the actual and former colonies/possessions of the British Empire. Common law has its *raison d'être* on precedents, praxis and routines of conduct rather than formal codifications of norms and principles. Common law is then systematising and ordering the customary practices in a more general and coherent form. Thus common law founds and unifies the law not by dictating an overarching structure from above, but rather by conforming and standardising the practice in a consistent way.

Bearing in mind this distinction, I would argue that the way GO performs its unification power, is closer, metaphorically speaking, to common law than to civil law. GO is unifying biological knowledge in a novel way that is different from theoretical unification but nevertheless practically useful and robust. Indeed, as the UK is a solid democracy without having a proper constitution, biology can be unified, in this sense, without having a general overarching theory. Indeed, as I have argued in the previous chapter, bio-ontologies were created with the precise idea that they could provide a unifying framework for the fragmented and variegated status of biological knowledge. In doing so, I have also shown that the epistemic strategy is to build a standardised and

computationally control vocabulary to address biological notions rather than creating an encompassing theory. Thus, a semantic solution. However, one may claim that if bio-ontologies are employing the role of theory in biology, then bio-ontologies are a sort of theory too although divergent from more traditional accounts. The problem seems not being pertaining to bio-ontologies, but rather to the, too narrow, views on what a theory is (or should be) in this type of contexts.

What are GO categories

Regarding this aspect some scholars (as Leonelli 2012) have precisely argued that forms of classification can count, in some cases, as specific forms of theory. According to Leonelli when the practice of classification is deemed to provide a common, shared and formal representation of the very elements being classified, this might lead to a something *theoretical* that, rather than imposed from the top, emerges from the data collection itself. As also claimed for taxonomical work (Müller-Wille 2007), in biology the collection principles are often embedded in the practice itself rather than lowered from abstract theories, thus framing from the bottom, the way biological 'facts' and 'objects' will be considered as such. According to this perspective, the type of generalisation provided by that kind of *classificatory theories* provides "an ideal of unity that only aims to establish some kind of commonality between different phenomena, without necessarily embedding that commonality within an overarching conceptual structure" (Leonelli 2012). Leonelli's reconstruction of ontologies unification power has definitely the merit to highlight how many scientific theories in the life sciences arose precisely from practice, thus broadening the very understanding of what may count as a theory. Since the role of such ways of

classification in shaping the research, by favouring the generation of new hypotheses and suggesting the type of experimental strategy, it should not be surprising to consider them a form of theory.

However, given the novelty of these tools and their divergence with other classificatory practices, I would rather propose to frame the specificity of ontologies in different terms. Indeed, if something is so different from the stereotypical idea of a theory maybe it is wiser to adopt a different perspective or to elaborate a new epistemic category. Thus, starting from Leonelli's endeavour, I go on by specifying how ontologies recall *models* in their architecture but remind *theories* in their behaviour. In order to do so, again is my focus is GO structure. By examining both the epistemic reasons for its implementation and the type of analysis provided by GO, I argue how such a tool resembles some features of a *model* but nevertheless constitutes something new in the epistemological scenario. Not entirely a theory, more than a model (however structurally similar to it), my point is that GO efforts constitute a novel category within the epistemic repertoire. Indeed, my central claim is the knowledge provided by GO, due to its regulatory objectivity, should be seen as a more or less effective tool through which we can discriminate, among an enormous amount of data, a *convenient* way of organising those empirical results which were at the basis of the GO analysis. In this perspective GO is a very peculiar map. A unique, *sui generis* instrument. An *orienteering tool* for biological research. Accordingly, such a specific status is better specified given that GO, as already argued, is both *conventional*, as the result of epistemic interactions towards a common agreement, and *normative*, since the tool shapes the representation of knowledge as it will be perceived by other, future researchers.

It is crucial then to specify first the features of such conventionalism and normativity.

In his famous *Science and Hypothesis*, Poincaré discusses about the epistemic nature of Euclidean axioms[17]. Poincaré famously argued that axioms of geometry "are merely disguised definitions" (Poincaré 1902). Accordingly, their nature and validity should not be established in terms of their truth or falsity. Rather, geometrical axioms "are *conventions*; our choice among all possible conventions is *guided* by experimental facts; but it remains *free* and is limited only by the necessity of avoiding all contradiction" (Poincaré 1902). Such a *conventionalist* position has an important consequence on the way we interpret representation. Thus scientific terminology does not semantically represent a state of affairs by mirroring a supposed metaphysical structure into a unifying definition. Science might have hidden ontological commitments but it is not directly interested in dealing with metaphysics. On the contrary, following van Fraassen (1980, 2008) I think that scientific terms aim to grasp and ordinate the sensible experience into a form that accounts for the empirical content but allows effective predictions and consistent explanations. That is to say that the language we adopt in science must be empirically adequate but it does not necessarily require a clear specification of a commitment on its truth. Such a perspective is definitively pragmatic and the debate on the nature of axioms in mathematics is certainly not over. However, I believe that this account of conventionalism works fine in our context. Indeed, moving to Gene Ontology, it is important to recall

[17] According to Kant, Greek geometry was 'true' and unique as constituting a privileged *a priori* form, shaping the sensible experience. However, the invention/discovery of non-Euclidean geometries had stimulated a heated debate about the truth of classical axioms. Thus the question was to clarify on what basis we should chose between a geometry and another one and which one is the 'real one'

again that such a tool has been developed to deal with knowledge management in the life sciences and not to solve metaphysical controversies on the nature of scientific categories (*i.e.* whether there are natural kind of not). Accordingly, I argue that the knowledge provided by Gene Ontology should not be evaluated in simply terms of its truth or falsity but rather as more or less effective tool through which scientists and researchers can discriminate, among an enormous amount of data, a *convenient* way of organising those empirical results which were at the basis of the GO representational analysis.

The view that GO is an *orienteering tool* means that it is an instrument through which scientist can *map* their data on a wider context and then, thanks to this, elaborate new experimental strategies. GO is truly a map for making the conceptual content of a particular experimental condition comparable across different research contexts. Such a map is essential not as a way to confirm experimental results but as a way to compare experimental results with the theoretical background (the so called 'big picture'). However the possibility of such a generalisation beyond the locality of data production does not create, *per se*, a unification for the theoretical content. Now, given that GO is a unification of "dis-unified fragmented research about a large variety of objects" (Leonelli 2012) it is important to clarify how exactly GO pursues such a task. Sabina Leonelli also claims that the generalisation provided by a tool like GO does not aim to universality but rather "only aims to establish some kind of commonality between different phenomena, without necessarily embedding that commonality within an overarching conceptual structure" (Leonelli 2012). Therefore GO unification power would be *reductive* (following Morrison 2007) and not *synthetic*. This is right if we attribute to synthesis a strong meaning (*a là* Hegel)

as the system arising from shattered parts and the resolution of a diversity at the same time. On the contrary, by adopting a different account, I will argue that GO provides indeed a *synthesis*. Let us see how.

By following Hacking (1983) natural phenomena are deeply related to the experimental dimension. Phenomena are not waiting for scientists to be discovered. It is the active enterprise of experimentation that generates the conditions by which phenomena become evident. Indeed phenomena are observable only under those specific conditions. To put it differently, it is the experiment that "creates" the phenomena. Such a claim must not be intended in an idealistic fashion but rather the opposite. In other words, phenomena are fragments of our understanding of nature and "in nature there is just complexity, which we are remarkably able to analyse. We do so […] by presenting, in the laboratory, pure, isolated phenomena" (Hacking 1983). Biology is particularly complex and very often experiments in biology do not point to general phenomena, but they rather isolate aspects of phenomena, due to the locality of biological data. Thus, the peculiar epistemic status of biology makes the *recomposition* of such a fragmented picture pretty hard. I suggest that GO fulfils this scope in a particular way. Indeed, unlike scientific theories, GO does not unify supposed (and still lacking) general biological principles but rather biological knowledge by creating a bridge among different *perspectives*. By that I link my argument to what Lorraine Daston (Daston 1992) calls *aperspectival* objectivity. According to Daston such an idea of objectivity started to be affirmed in the late 19th Century as the type of scientific objectivity *par excellence*. This notion of objectivity was conceived as opposed to subjective idiosyncrasies and interpretations that might occur in scientists' mind. Thus, in

this context, ‘aperspectival’ means precisely an absolute point of view, or as famously Thomas Nagel argued, “a view from nowhere” (Nagel 1989). As Daston writes “[a]perspectival objectivity was the ethos of the interchangeable and therefore featureless observer - unmarked by nationality, by sensory dullness or acuity, by training or tradition; by quirky apparatus, by colourful writing style, or by any other idiosyncrasy that might interfere with the communication, comparison and accumulation of results” (Daston 1992). My argument rests precisely on the idea that experiments are, on the contrary, always *perspectival* in an epistemological sense. By that I mean that, although methodologically exemplary, experiments reveal *aspects* of the world. There is no natural science of the ‘absolute’. I try to clarify myself by an analogy with visual perception. In cognitive understanding, and more importantly, in pictorial representations as paintings, objects are always given to perception according to a particular perspective. I do not perceive the book on my table in its wholeness. Cognitive knowledge always presents *perspectives*. Within the lab, the model, the experimental setting, material and methods are the counterpart of visual perspectives. As I recognize the book on my table as a whole through the composition of different points of view, also the knowledge about biological phenomena derives from many different experiments. The phenomenon is then the idealised recomposition of these perspectives. Let us consider a simplified example. A type of experimental research is aimed at determining the transcription factors binding site within a specific cell type (*e.g.* macrophages), in specific conditions (*e.g.* after LPS stimulus). A second experimental strategy might analyse the same conditions but with an eye on chromatin structure. A third experimental team would do the same but it would change some conditions

as the cell model or the type of stimuli. A fourth approach would ignore the expression profiles and will focus on some functional products associated with those genes (*e.g.* proteins) trying to determine their structure. In the end, a fifth lab may be interested in a proteomic analysis of those functional products. All these approaches are pointing at biological phenomena but they could not grasp them in their entirety. This, again, does not conflict with the idea of a possible objective knowledge. Indeed, such a picture fits well with the conception of scientific endeavour invoked by the so-called *scientific perspectivism*. "Overlapping perspectives, whether observational or theoretical, suggest that there is 'something' there" (Callebaut 2012, p 76). However knowledge is precisely the interplay between different perspectives rather than the absence of a perspective. Experimental results are thus *perspectival representations* on phenomena. These perspectives represent precise *directional understanding* about more general phenomena. Data coming from these experiments are then more abstracted and crystallised via models. An example of this kind of model could be the scheme of particular molecular pathway or the sketch of a proposed mechanism of action of a protein. As coming from different sources these data are dispersed in the fabric of knowledge. A map can help to locate them. This what GO can actually do. As a cubist painting, ontologies are 'synthesising' models of data into meta-models about more general phenomena. Yet, unlike cubism, such a synthesis must not be intended as a mere composition of perspectives. Indeed, one might see the notion of synthesis as the conflation of different aspects into something new that resembles all of them. If this is the case, GO would create the view from nowhere. However, in this case GO would not fulfil its purpose, that is to make possible to situate and retrieve the

information and the knowledge displayed in the big picture within the original context. Such a confusion can depend on the account of synthesis we adopt. That is why such a synthesis should not be understood in a simplistic way. The kind of unification provided by ontologies does built neither an absolute nor a privileged vision as it were a view from nowhere or from God's eye. The integration granted by ontologies is neither grasping nor highlighting any supposed, *essence* of biological phenomena. Ontologies are then a unification tool in the sense that they construct a grammar of translation, a *dictionary*, according to which scientists can pass through the 'different languages' by which biological phenomena are given to our knowledge. In other words, ontologies are not conflating different perspectives into a new, single, unified and complete point of view. They rather provide a common, semantic framework on which diverse epistemic contents can be compared to each other into the whole fabric of biological knowledge. Thus Leonelli is definitely right when she claims that GO generalisations are not an *overarching synthesis*. However here lies the core of the question. I think that Leonelli's position, and her rejection of the idea that ontologies can provide a synthesis, is imputable to the image according to which principle guiding bio-ontologies are thought. Indeed, as I have shown, in constructing ontologies, many computer scientists based their formal work on the analytic framework based on a particular way of thinking that has been derived from the interpretation of Aristotle's work (Smith 2007, Smith and Ceusters 2010) meaning that most of current bio-ontologies are built on, or at least inspired by, an upper-level ontology called BFO (Smith 1998, Smith and Ceusters 2010)[18]. While such a structure can work perfectly fine within an

[18] Just to recall, BFO is "designed for use in supporting information retrieval, analysis

internal viewpoint, (*i.e*. the research procedures and practical uses of bio-ontologies) in the light of a more refined epistemic analysis many problems arise.

In particular, it is fundamental to clarify in what sense a term in GO aims to provide a *general form* (the semantic definition) unifying different perspectives about a phenomenon. In other words, since behind any GO term there is a related concept, notion, piece of biological knowledge represented and condensed by that term, the question is to furnish an analysis of such concept formation happens. In doing so, I briefly recall the classic theory of concept formation (elaborated by Aristotle and his followers) and then its critique developed by Ernst Cassirer (Cassirer 1910). Then I present a view according to which GO terms should not be based on the classical theory. Last, I apply such a different approach to GO terms showing how this better mirrors the use of GO in the practice of science.

In his theory of concept formation Aristotle does not specify a clear difference between the *object* and the *concept* which is, in an epistemic setting, fundamental. But more importantly, according to the Greek philosopher "defining a *term* means giving the essence of what it refers to, that is, it means giving the 'what it is', or determining the species to which what is referred belongs. For to define a term is sufficient to determine the proximal genus and the specific difference which distinguishes that species from the other species of the same genus. While the genus determines the essence in an undetermined way, the difference focalizes the essence by indicating the species" (Boniolo

---

and integration in scientific and other domains. As mentioned before, BFO is an upper level ontology. Thus it does not contain physical, chemical, biological or other terms which would properly fall within the coverage domains of the special sciences. As already shown, BFO is framed into a sort of a simplified Aristotelian scheme, in which information content is fundamentally described in terms of *substances* which possess properties or *accidents* (Smith 1998).

2007, italics is mine). Therefore, following Cassirer, for Aristotle the individuation of a shared commonality is pursued via *abstraction*. What makes a horse as such, so what all horses have in common, is identified by abstracting the form from the accidental properties of single occurrences. According to Cassirer's analysis, the theory of Aristotle must be rejected since it represents concepts as *containers*. On the contrary Cassirer argues that, "the concept should no longer be understood as a class containing objects, but as a representation unifying given objects in a given way by giving them […] significance. Consequently the *form of a concept* has a fundamental and primary role: to rule the intellectual synthesis" (Boniolo 2007). Cassirer's view is notoriously embedded in Kantian tradition. Kant nowadays is a bit 'out of date' in philosophy of science. Nevertheless his Copernican Revolution reminds us the importance of the distinction between ontology and epistemology. A cornerstone after which, as philosophers, we should not go back to. Thus, it is not sufficient just to observe in order to grasp similarities among natural phenomena. It is rather the way by which scientists deal with the objects of their interest, that builds such a similarity.

Let us shift from theoretical rarefaction to a more concrete situation as bio-ontologies. If we examine any GO term, we could appreciate that they are not construed by the mere accumulation of data. The *semantic synthesis* of experimental evidences pursued by GO is a practice that requires the term being regulating the future use of the term itself. Following a Kantian vocabulary, in our case of interest, the GO term provides the 'unity of rules' for the use of the term itself. An important difference with Kant and Cassirer here, lies in the fact that the level of such an operation is not just the single epistemic agent but the

entire scientific community.

Giovanni Boniolo has already explored such issues in relation to the semantic of standard scientific terminology. “What does possessing a rule mean? Nothing but knowing how to apply it and, […] knowing what satisfies it. […] In other words, grasping the rule means grasping the concept, more precisely its sense” (Boniolo 2007). Following Boniolo’s analysis, I define with *semanticizing area* the set of rule required in order to form a concept. In this perspective, the semantic sense of a concept is not something inherent to the concept itself but rather to the ways according to which such a concept has been synthesized. I already described that the main aim of GO is to build a unified vocabulary for scientific terminology within molecular biology. However, as also already shown, a great issue in achieving such a task is the semantic divergence among scientists that reflects the differences (aims, model organisms, methodologies) present in diverse fields and subfields of the research. Thus, one may interpret GO as a way to systematise, in an ordered manner, the diverse *semanticizing areas* which contribute to grasp the meaning of interested, relevant phenomena. By adopting such a view I argue that the presence of *semantic fluctuations* in GO terms, due the advance of scientific research, are explicable as variations of the semanticizing areas concerning those specific phenomena under investigation.

To sum up in this chapter I tried to show how and why a tool like GO is a kind of epistemic map, an *orienteering tool* for biomedical research. Moreover I also presented an argument to clarify how GO would unify biological knowledge and what is the meaning of such a unification.

Now it is time to look at the real practice. In the next chapter I will examine several specific scientific cases, such as a recent article published on Nature Genetics, that heavily relies on GO.

CHAPTER IV

Nel suo profondo vidi che s'interna,
legato con amore in un volume,
ciò che per l'universo si squaderna:
sustanze e accidenti e lor costume
quasi conflati insieme, per tal modo
che ciò ch'i' dico è un semplice lume.

Dante, Paradiso, Canto XXXIII, vv. 85-90

Doing science with GO and beyond

In this chapter I will analyse how GO constitutes a driving force for contemporary biomedical research. In doing so, along with the brief examination of some cases, I will focus on a particular example, a recent article, "*7q11.23 dosage-dependent dysregulation in human pluripotent stem cells affects transcriptional programs in disease-relevant lineages"* published on *Nature Genetics* (Adamo, Atashpaz, Germain *et al.* 2015). The article provides an excellent, paradigmatic, case of how a tool like GO changed the practice of current scientific research in the life sciences. Indeed, in the paper, the experimental strategy, the methods and the rationale are all embedded in a map thinking framework. However there is more. Such an article really shows a peculiar and distinct, way of doing science. This way is not simply ascribable to naïve dichotomy between hypothesis-driven vs. data driven. It is rather a complex combination of the two, in which the knowledge coming from databases, exploited by a tool like GO, drives the other components of scientific

efforts by exchanging the epistemic primacy and priority of exploratory experiments with the navigation in the data sea.

The aim of the article is ambitious. Summing up, its purpose is to increase the reliability of *induced pluripotent stem cells* (iPSCs) as models for diseases. iPSCs are a type of cells, undergone molecular reprogramming, presenting, *bona fide*, the features of embryonic stem cells. iPSCs constitute a promising and trendy sector of biomedical research since they challenged the idea that specialised cells are inescapably committed to their fate. In 1960s John Gurdon (Gurdon, 1962) has shown how somatic, differentiated cells could be turned back into their embryonic state by transferring nuclei of epithelial cells into enucleated eggs of a frog. Recently, Shinya Yamanaka (Takahashi and Yamanaka, 2006) has demonstrated that, without any nuclear transfer, he could reproduce Gurdon's results by exposing differentiated cells to specific factors which eventually turned (*i.e.* reprogrammed) those committed cells back into their pluripotent state (*i.e*. iPSCs). Such a discovery granted the Nobel Prize in 2012 to both of them, and opened the door to the implementation of iPSCs in many areas of biomedical research. One application is precisely *disease modelling*. The potential of such an approach is that iPSCs should allow a better analysis of the complex picture of pathogenic drives in a developmental context via molecular approaches. In other words iPSCs could shorten the gap between clinical and research contexts by permitting the track of the consequences of genetic alterations through the cell development thus providing hints of clinical relevance. In order to exploit such a potential it is fundamental to provide an answer to, at least, two problems. On the one hand it is central to determine how much genetic alterations, in early developmental phases, are indicative over

related pathological conditions and their molecular pathways (*i.e.* to observe the onset of the disease in a preclinical phase otherwise not detectable). On the other hand it is crucial to establish how much iPSCs modelling is apt to identify these pathways. Moreover, this approach to disease modelling could, in theory, provide suggestions on relevant molecular mechanisms from the point of view of future therapeutic implementations.

The authors of the article addressed then these issues by investigating two related genetic syndromes produced "by symmetrical copy number variations (CNVs) at 7q11.23[19] involving, respectively, the loss and gain of 26-28 genes: Williams-Beuren syndrome (WBS) and Williams-Beuren region duplication syndrome […] that includes autistic spectrum disorder (7dupASD)" (Adamo, Atashpaz, Germain *et al.* 2015, p132). The main idea is that, thanks to iPSCs modelling, it would be possible to scrutinise such a biological symmetry (let us consider that genomic inverted alterations are mirrored by specular, behavioural phenotypes) starting from the 'origin' or the stem like state. Thus, due to a collaboration with clinicians, researchers were able to have samples from a cohort of patients resulting in 4 different genotypes: the WSB typical deletion, WSB atypical deletion (a shorter one, in terms of base pairs, and less frequent), the control case and the 7dupASD duplication. Skin fibroblasts have been reprogrammed via synthetic mRNA encoding different pluripotent factors, thus developing a total of 27 iPSC lines. Successively, the pluripotent state of these cells has been confirmed through transcriptomic analysis[20]. Such a step is a further confirmation of the standardising power of databases for research. Indeed, the pluripotent state has been determined as such since the transcriptomic

[19] a genomic region on human chromosome 7
[20] and also by IF (immunofluorescence) of pluripotent factors

profile has been compared and matched with published datasets. RNA-seq (roughly, the sequence of the transcription) and Nanostring quantification (another methodology to assess gene expression) also confirmed that gene expression mirrored gene dosage and, again via *database consultation*, scientists were able to verify that two proteins, GTF2I and BAZ1B, are "encoded by genes associated with key traits of WBS and 7dupASD" (Adamo, Atashpaz, Germain *et al.* 2015, p133). In particular, GTF2I protein level correlates with gene dosage. Next, differential expression analysis between distinct genotypes has been conducted by RNA-seq profiling of iPSCs, then comparing the results against some control cell lines. This is again through the use of databases, that provided the reference context on which to give sense to experimental results. A pairwise comparison of the three genotypes (Williams-Beuren syndrome vs. Control Group, Williams-Beuren syndrome vs. 7dupASD and 7dupASD vs. Control Group) revealed 757 *differentially expressed genes* (DEGs). Finally, a GO term enrichment analysis of the union of DEGs has been performed.

At this stage Gene Ontology comes explicitly into play. However, I believe that GO rationale has driven much part of the experimental design and has highly influenced the earlier steps of such a research study. In other words my argument is that GO is at the basis of the main heuristic strategy of the entire study. In order to justify my claim, before discussing the use of GO and its results, I will go back to the previous phases of the experimental strategy in order to detect and unveil the role of GO.

Above all, there is the combined use of iPSCs and Gene Ontology. As already mentioned, iPSCs constitute an excellent surrogate of embryonic stem cells in terms of pluripotency and stem like features. Indeed, whereas the process

of reprogramming had been conducted effectively, it would be virtually impossible to distinguish iPSCs from ESCs (embryonic stem cells). *Pluripotency* is defined as the capacity of a cell to differentiate itself into any other cell of an organism (see for instance the Oxford Dictionary of Biology, 6th Ed., 2008). At a molecular level, cell types are determined by peculiar gene network interactions. Being pluripotent means thus that a cell shows a particular molecular signature established by the modality of genes' activity. Indeed, since the genome of any cell of an organism is almost the same (there are some exceptions, but it is not fundamental to discuss this point here), the differences among cell types and states should be mainly attributed to the way genes are differentially expressed and regulated. Thus the pluripotent state (as any other cell state) is essentially related to epigenetics, or how different parts of the genome are alternatively transcribed, silenced and modulated.

In the case discussed here, the creation of iPSCs lines from patients affected by WBS and 7dupASD syndromes, can potentially allow scientists to obtain specific cell types (such as neurons) for further experiments. More than a quip, this could mean that it would be possible, in theory, to have a "brain in a dish" (see for instance Shen 2013). However, the authors of the article do not pursue that path (although it is possible that they will do in the future). Why is it so?

First, the production of specific cell differentiated lines is not straightforward. Both reprogramming and transdifferentiation (the artificial induction of a somatic cell to commit itself to another cell state) are not an easy task to perform. Due to technical difficulties some cell types are either almost impossible to obtain or the efficiency of the procedure is so low that becomes useless (see for instance Hanna, Saha and Jaenisch 2010). Second, the materiality

of somatic cell lines does not ground, *per se*, a better explanatory framework. Indeed cell cultures do not constitute a reliable model in virtue of simple similarity. Moreover, given the complex nature of both syndromes, it would be very difficult to assess which neurons (among different types) will play a role, and how they do so, in the disease. In addition, it would also very troublesome to reproduce the material structure of relations of a brain, just through neuronal cultures. However, as explained in the previous chapter, this does not prevent a thing such as a cell from being a good model. Every scientist is aware that a bunch of cells does not properly portray all the features the related tissue and organs. Still, as also shown in the previous chapter, this does not prevent a thing such as a cell from being a good model. But a good model for what?

The choice to focus on iPSCs is motivated precisely because they can provide a better model, compared to cultures of differentiated cells, for developmental conditions, given the implementation of certain type of analysis. It is also, again, a matter of style. Indeed, the type of evidences obtained through empirical experimentation are epistemically different from those coming from computational approaches. Certainly, the material production of distinctive cell types (*i.e.* neurons) is not mutually exclusive with bioinformatics work. On the contrary, they are complementary. However different research strategies will prefer some kind of evidence over other types of it. In our example, the production of specific cell lines, with no other indication, could have been potentially uninformative. On the contrary, the adoption of a tool like GO, allows to global map a sort of 'cell differentiation process *in silico*', thus suggesting what to look at in further experiments. This is possible because of the combination of the features of iPSCs and ontologies. As already said, within

iPSCs there is all the potential of developing every cell of the organism. This means that iPSCs, given the genetic nature of the diseases taken into account, can contain, virtually and *ab origo*, all the relevant elements that could affect the molecular phenotype of interest (and hopingly suggesting therapeutic interventions in the clinical setting). On the other hand GO, being an updated, global map of biological knowledge, allows to compare local findings with those ones coming from other experimental settings and to situate them into a wider picture. GO permits then to computationally explore the space of possible relations of different cell lineages through the comparison of given samples against all the relevant data stored in databases. Therefore, the possibility granted by GO, shapes indeed the type of scientific strategy. A strategy that can be described as: first, making a map.

Let us examine the reason why the construction of such a map is possible and probably, needed. First the kind of data. This type of science is heavily based on *omics*. Normally, transcriptional profiles show the global set of all RNA transcripts of a given genome (under specific conditions). However, their analysis affect cell populations rather than single cells. As a matter of fact, no cell behaves exactly as others, even if it is of the same type, in the same context. Uniqueness and intrinsic variation are indeed features of biological objects since 19th Century natural history. This means that, by performing transcriptomic analysis, scientists normally privilege the understanding of the average behaviour rather than the detailed (hopefully mechanistic) description of single cell behaviour and its fluctuations and relation with the other cells[21]. Transcriptional

[21] Some researchers have argued the necessity to improve single cell analysis study. This is perfectly fine and compatible with what I said, as it will also respond to different epistemic desiderata. See again for instance Hanna, Saha and Jaenisch 2010.

profiles are indeed general, cell-group behavioural maps. Certainly a map of this kind misses something. Tiny differences will be neglected and 'absorbed' by the background. This is not a problem. As I said in the first chapter, a map as much detailed as the object it represents, is basically useless. Indeed, when the authors of the paper have identified 757 DEGs they did not care (for that moment) about *how* (i.e. which mechanism was responsible for it) these genes were differentially regulated (probably many genes are altered in different ways, one from another). Their concern was about *where* this distinct regulation happened. By that I mean that scientists have looked at the "number and distribution of DEGs across the comparison among the three genotypes" (Adamo, Atashpaz, Germain *et al.* 2015, p134, fig 2a) rather than looking at the mechanistic nature of such regulations.

Thus, these finding are suitable precisely to build the kind of map in question. If iPSCs 'contain' the entire horizon of developmental possibility (in terms of different cell types) and their transcriptional behaviour suggests the directions of such a development, GO is then a map to navigate this computational horizon. And GO allows such a 'virtual tour', not by virtue of a direct and experimental examination of specific cell type lines, but rather through the fact that this tool is capable of computationally disclosing the information that is biologically enclosed in pluripotency. Indeed, as the authors themselves comment "[s]trikingly, Gene Ontology (GO) analysis of the union of DEGs showed significant enrichments for biological processes of obvious relevance to the hallmark phenotypes and target organ systems of the two conditions" (Adamo, Atashpaz, Germain *et al.* 2015, p134). This shows very well why GO is an orienteering tool. GO is able to situate the information coming from the experimental work into the most updated map of current biological knowledge,

thus highlighting connections and relations practically invisible to any single researcher or group. The power of GO is therefore to unveil existing, hidden links of biological knowledge. If the map of species and organisms provided by taxonomists is capable of suggesting possible indications on the relationships among those species, then the map provided by GO shows the capacity to do something similar for biological processes at molecular level. GO thus revealed that "[t]he top-ranking categories were related on one hand to cell adhesion, migration and motility, which appear especially relevant in light of the wide range of connective tissue alterations that characterize WBS, and on the other hand to the nervous system, providing a molecular context for the defining neurodevelopmental features of the two conditions. Additionally, further enrichments were related to remarkably specific features of the two diseases, including (i) cellular calcium ion homeostasis, a category of potential relevance across disease areas but that acquires particular salience given the high prevalence of hypercalcemia in WBS; (ii) inner ear morphogenesis, consistent with the hyperacusis and sensorineural hearing loss in WBS, as well as with the balance and sensory processing alterations found in ASD; (iii) a number of categories relevant for the craniofacial phenotypes, as represented by several categories, such as skeletal muscle organ development, migration and neural crest cell differentiation; (iv) blood vessel development and cardiovascular system development, reflecting the wide range of cardiovascular problems in WBS; and (v) kidney epithelium development, in line with the highly prevalent kidney abnormalities in WBS" (Adamo, Atashpaz, Germain *et al.* 2015, p134).

The possibility of such an approach suggested also a further step. If the GO analysis on iPSCs provided such a global map, the researchers, in order to prove

whether transcriptional dysregulation would be amplified during development, derived also three lineages of cell types precursors: *PAX6-positive telencephalic neural progenitor cells* (NPCs, responsible for radial glia cells formation which, in turn, form cerebral cortex); *neural crest stem cells* (NSCSs, involved in the formation of craniofacial structures; and *mesenchymal stem cells* (MSCs, which are progenitors of osteocytes, chondrocytes and other cell types relevant for both syndromes). All these three lineages are crucially significant for the pathological conditions under examination. The GO analysis can be seen here as the creation of three sub-maps of the previous one, against which they should be compared and judged. Indeed the researchers "evaluated, for each of the three differentiated lineages under study, the proportion of DEGs showing conservation of the GO categories that were found to be enriched in iPSCs. Upon differentiation, iPSC DEGs were preferentially retained by category in a lineage-appropriate manner such that, for each target lineage, the proportion of conserved iPSC DEGs was much greater in categories relevant to that lineage (such as axonogenesis and axon guidance in the neural lineage, synapse-related categories in NCSCs that originate the peripheral nervous system and smooth muscle–related categories in MSCs)" (Adamo, Atashpaz, Germain *et al.* 2015, p138).

By looking at the conclusion of the study, it is quite clear that the main result of the research is the production of a specific kind of map. In particular, GO served perfectly the purpose of exploiting the potential of iPSCs. First, GO was an indispensable tool in order to manage the intrinsic *variability* of iPSCs as model for diseases, given that such a variability occurs across both individuals and lines derived by the same individual. Indeed, in order to obtain a reliable, and as much as global, picture, variability has had to be taken into account and

produced, by the creation of the greatest cohort of iPSCs lines for any relevant condition. Next, of course, all this information should has been processed via high-throughput approaches. As in a complex forecast model where scientists need to take into account and compare different kind of data such as geographical details, temperature differences, winds' directions and intensity, geological factors etc. and to display all of them on a common representation format, here the different transcriptional behaviours of distinct genetic conditions, the developmental issues and the pathological considerations were all consistently represented and managed by GO analysis. Indeed, such an approach perfectly exploits the potentiality embedded in iPSCs by predicting, already in the pluripotent state, which pathways will be affected given the specificity of the conditions under investigation. Moreover, the creation of a such a map, is indeed an orienteering tool by which scientists navigated the developmental trajectories thus showing how such a dysregulation "selectively amplified in a lineage-specific manner, with disease-relevant pathways preferentially and progressively more affected in differentiated lineages matching specific disease domains" (Adamo, Atashpaz, Germain *et al.* 2015, p139).

Once such a complex, multi-map has been built, it is also possible to better locate and address single factors (such as that particular protein) into the wider context of the disease development, thus suggesting further possible steps and experimental approaches. Indeed, as the relevance of specific gene products is globally assessed by GO analysis, then it would be possible to better focus on them (also with more traditional, mechanistic approaches). As the authors themselves argue "[n]otably, our analysis of symmetrically dysregulated targets also uncovered the following genes as prime candidates for mediating the

molecular pathogenesis of defining aspects of the two conditions: (i) *PDLIM1*, which has been associated with ADHD, neurite outgrowth, cardiovascular defects and hyperacusis; (ii) *MYH14*, which is involved in hearing impairment; and (iii) *BEND4*, encoding a transcription factor harboring the BEN domain that distinguishes a recently characterized family of neural repressors and that was sensitive to both *GTF2I* dosage and its LSD1-mediated repressive activity, a finding that also resonates with the inversely correlated pattern of *GTF2I* and *BEND4* expression in the human brain" (Adamo, Atashpaz, Germain *et al.* 2015, p140). Such a scientific contribution does not certainly exhaust all the possibilities of map generation in this context. On the contrary it promotes the implementation of new maps and it suggests possible directions for more traditional, mechanistic experiments in order to investigate the single elements displayed on the generated map. As argued before, such efforts will be better addressed given the standardisation created by GO. Hence, in order to promote and enhance such a common frame, researchers have also designed a web platform, named *WikiWilliams-7qGeneBase* to make data available to the research community working on these syndromes. Such a database will be open to external contributions given the adherence to shared format principles. In the end, by granting an original kind of scientific results, aimed at disentangling some crucial aspects of complex syndromes, and by contributing to the implementation of regulatory standardisation procedures in data display and management, such a study provides a clear example of a new way to conceptually and experimentally address the practice of epigenetic studies and transcriptional analysis.

Although quite innovative, such an example is not isolated. As argued in a recent publication (Hoehndorf *et al*. 2014) the use and the importance of ontologies in biomedical research have radically increased. This is due both to the amount and the type of data produced in many areas of biological research. Ontologies rapidly became a key tool in the interpretation of data and they fostered the creation of biomedical IT infrastructure, such as the Elixir initiative (http://www.elixir-europe.org). The purpose of ELIXIR is to coordinate the collection, quality check and archiving of large amounts of biological data generated by all varieties of experiments in the life sciences. Some of these datasets are extremely specialised and would previously only have been available only to those researchers working in the context in which they were produced. Ontologies are thus becoming, being already a fundamental instrument for the interpretation of data, an inspirational resource for users and developers in building new analysis methods.

In this sense, a good example is constituted by a recent research (Yang, Chen *et al.* 2014) that adopts GO terms in order to highlight probable common features of tumour suppressor genes (TSGs). The aim of that kind of research is precisely to enhance the creation of building effective prediction methods for the identification of TSGs. The idea is that if it is possible to individuate a set of properties shared to all TSGs and express it in a semantic representation framework, then such an information could be extremely useful in order to detect and discover new TSGs. In other words, scientists have first adopted a huge database known as *TSGene* (http://bioinfo.mc.vanderbilt.edu/TSGene/) which is a repository of known, validated tumour suppressor genes. Next, researchers used GO terms enrichment analysis as parameters to encode the genes of interest

within the TSGene database. GO terms have been easily selected as codes for TSGs since the database itself has been constructed in a GO friendly way. The first analysis on GO *biological process* domain, ranked top five terms: GO: 0022610: biological adhesion, GO: 0040007: growth, GO: 0032502: developmental process, GO:0065007: biological regulation and GO:0050896: response to stimulus. Both *biological adhesion* and *response to stimulus* resume important features of TS proteins as they are involved in the alarm reaction and they show a guardian role both in tumorigenesis and in the metastasis formation. The term *single-organism process* has also been highlighted. This again suggests and confirms a very known property of TSGs, since the fate of an organism dramatically depends on the cell cycle and apoptotic processes. Indeed, TSGs play a crucial role in the conservation of the cell cycle checkpoints and in the induction of apoptosis. The same approach has been used for *cellular component* and *molecular function*. In the first case the most important terms are GO: 0030054: cell junction and GO: 0044422: organelle part. Again the first term reflects the very well known fact that some TSGs are as such because of their relevance for cell adhesion which is a key component of metastatic process. While the second term confirms the involvement of organelles (such as mitochondria, ribosomes and ubiquitin-proteasome system) in the phenomenon of tumorigenesis. Lastly, concerning molecular function the most represented terms are GO: 0005488: binding, GO: 0003824: catalytic activity, GO: 0030234: enzyme regulator activity, GO:0004872: receptor activity and GO:0060089: molecular transducer activity. Also in this case GO terms individuate properties which are considered crucial for TSGs. Given the type of results, a biologist could ask why making such an effort if the results are, in a sense, already known.

This is certainly true but such a picture does tell just the half of the story. The map produced by GO is a first attempt to have all the properties shared by TSGs addressed in a common format. This means precisely that functions, cellular structures and activities can be all taken into account in order to construct a common signature for the features defining TSGs. Thus, such a knowledge, given also the way it is displayed, can definitely constitute the first step for future heuristic strategies aimed at the discovery of new TSGs. Indeed the researchers have tried "to predict the novel TSGs based on features in the total optimal feature set, *i.e.*, the key functions that defines tumor suppressor. For each 'negative gene', we counted the number of key tumor suppressor functions that it was annotated onto. The genes with great number of key tumor suppressor functions were considered as candidate tumor suppressors, since they shared similar functions with the known tumor suppressors" (Yang, Chen *et al.* 2014, p 9). Finally this analysis revealed a list of possible candidates (based on the fact that these genes share 293 annotations with known tumour suppressors) that will be proposed for further experimental and clinical validation. As argued in the previous chapter, GO is not a discovery tool *per se*, but it is an enhancer for discovery strategies. It is a map to navigate biological knowledge.

A third example (Cheng *et al.* 2014) affects directly areas of research that traditionally have been addressed by mechanistic approaches such as gene function prediction. In this case GO analysis is adopted to extract the information on function dispersed within the biomedical literature in order to enhance gene function prediction. From a methodological perspective, such a task can be achieved due to the hierarchical structure provided by GO. In other words, researchers first adopt a literature-based method (via text-mining tools)

extracting gene function from annotations. However such approaches, in the past, have shown a low accuracy in function prediction. One possible explanation is that previous attempts ignored the hierarchical structure of GO. Indeed, being a map, GO does not simply represent the content of biological knowledge but it also display how such a knowledge is structured and hierarchically organised. This aspect shows very well how GO is precisely representing biological knowledge rather than hypothetical natural kinds. It does not stand for what is out there, but it illustrates our epistemic categories in the analysis of natural world. Moreover, biologists know very well that genes may likely have more than one function, depending on the environmental and temporal context, the type of cell considered, the tissue to which those cells belong and the expression of other genes. The same gene product can then been subjected to multiple annotations which, in turn, contribute to different terms. Therefore, if one adopts a traditional classificatory approach, it would result that one instance should correspond to one class, while in biological reality it may pertain to many classes. Because of that, researchers have implemented a multi-label classification in which each gene is likely to be associated with several GO concepts. In other words, "[f]or example, the gene P25686 in the UniProt database is annotated by GO terms with *id* 0032436 (positive regulation of proteasomal the ubiquitin-dependent protein, catabolic process), 0090086 (negative regulation of protein deubiquitination), 0030433 (ER-associated protein catabolic process), 0031398 (positive regulation of protein ubiquitination), 0090084 (negative regulation of the inclusion of the body assembly). These *GO concepts together describe the gene functions*: protease-based pan-hormone catabolic process positive regulation of protein de-ubiquitin

negative adjustment, the ER-associated protein catabolic process, positive regulation of protein ubiquitin, the virus endosome assembly negative regulation. Therefore, we may regard the prediction of gene function as a problem of multi-label annotation, namely selecting several GO concepts as function description of a given gene" (Cheng *et al.* 2014, p 2, emphasis is mine). This study offers a clear example of how this new way of doing proceeds. Rather than taking one gene and empirically check its putative functions by implementing different experimental conditions, such a result has been obtained by a tool capable of navigating different databases. In other words, while the traditional approach started from the object (the gene) in order to investigate certain properties (functions) of it, this new approaches begins the other way round, with mapping all the known functions and then trying to attribute them to single genes due to the power of big data.

The rise of ontologies as drivers for scientific research does not stop with GO. As already argued, in the last years many other ontologies have been developed and implemented (consider for instance the aforementioned *OBO Foundry* initiative), sometimes orthogonally integrated with GO itself. It is the case, just to mention some examples, of Sequence Ontology (SO, http://www.sequenceontology.org/) that is aimed at the management of the increased production of more and more sequencing data in order to built a compatibility framework for the characteristics of diverse data formats of genomic sequences. On the other side, we have Protein Ontology (PRO, http://pir.georgetown.edu/pro/), the first logically-based classification of diverse classes of proteins. By gaining information from different types of databases, PRO representations include protein isoforms and variant, naturally and

artificially modified forms (due to biotech innovations) and also protein complexes. In order to do so, PRO is articulated in three sub-ontologies, *ProEvo* which classifies proteins according to their evolutionary relatedness, *ProForm* grouping proteins assembled by a specific genetic locus and *ProComp* which deals with specific amino acid chains presenting complexes. Another good example is constituted by Celltype Ontology (CL, http://www.obofoundry.org/) whose purpose is to construct a formal representation of cellular phenotypes among different organisms.

The proliferation of ontologies in biomedical research represents not only a new sign of scientific creativity but, as it is part of a collective and organised enterprise (involving consortia and diverse groups), it also responds to specific needs of the scientific community. Thus such an explosion provides an indication of the transition from the descriptive side to the normative one exposed in the previous chapter. Moreover, the limitations of a given ontology can be overcome by its implementation with other, orthogonal resources. For instance some researches have recently suggested (Wittkop *et al.* 2013) that term enrichment analysis based on GO, although widely used, shows the best efficacy on predefined gene annotations that are restricted to those domains highly manually curated. In order to cope with this issue, by developing tools hybridising the analysis of manually curation with automatic text search from other ontologies, it would be possible to expand the set of hypotheses created by term enrichment analysis. In this respect it is interesting to examine a case (*i.e.* Washington and Haendel *et al*. 2009) in which scientists adopt several ontological resources to deal with more traditional problems. That is to say that here researchers do not just produce different types of evidences and procedures, neither they just

endorse new perspectives and angles through which they could look at their object of investigation. In this situation they also tackle a very classical problem by addressing it with a new *way of doing*. I will further examine the features of this approach and its philosophical implication in the next section. For the moment, let us just delineate the main characteristics of such an effort. The scope of the article is to create a formal, shared, frame in order to better ground the connection between the experimental results obtained via animal models with the human diseases. The relation between the 'bench and the bedside' (*i.e*. how to exploit the findings of scientific research in order to develop new drugs, tools, and therapeutic interventions at the clinical level) is quite problematic (also in its definition) and constitutes an entire area of research called *translational medicine* (see for instance Woolf 2008). One of the main issues at stake is that phenotypic outcomes of mutations are generally based on criteria pertaining to different organism-communities and therefore they are described according to specific semantic choices, often anchored to the anatomical and physiological peculiarities of the animal models under consideration. Moreover, while the research on model organisms is often centralised and already subject to forms of internal standardisation, human-focused biomedical research does not present a clear and established form of commonality in terms of database structures and resources (see also Leonelli 2012). In addition, despite the presence of methods for comparing sequences (such as BLAST algorithm) the genetic basis of many diseases is still obscure and most the clinical classification of diseases rests on phenotypic descriptions. Thus, the main idea of using ontologies, is to have a tool to standardise and compare phenotypic descriptions across species and databases. Let us briefly examine the strategy adopted by Washington, Haendel

and colleagues. First, they grounded their analysis on a general database of phenotypic resources: the *Online Mendelian Inheritance in Man* or OMIM. Second, each phenotypic character has been registered through the combination of two elements (named EQ method): an *entity* (E, such an anatomical part or a process) that bears a *quality* (Q, such as big, increased temperature etc.). Interestingly such a classification procedure is the result of the merge of terms coming from different ontologies. Indeed, while entities are usually extracted by GO or other anatomical ontology, qualities come from PATO (Phenotype and Trait Ontology) which is orthogonal and fully integrated with GO and other OBO ontologies. "For instance, a *Drosophila* 'redness of eye' phenotype could be described using the terms "red" from PATO and "eye" from the Fly Anatomy ontology (FBbt) into the EQ statement EQ = `FBbt:eye` + `PATO:red`." (Washington and Haendel *et al*. 2009, p3). Next, researchers have tested whether EQ system can ease and reveal possible relations between genotypes and phenotypes across different species. Then, they used the EQ classification method to annotate 11 human disease genes from OMIM database to create a sub-dataset suitable for cross-species comparison. In order to cope with different anatomical structures pertaining to different organisms (zebra fish and mouse in this case), scientists have also developed a cross-species unifying ontology for anatomical structures (UBERON, http://uberon.github.io/). In the end, researchers were able to show that, within the same organism but also in different species, most of the allelic variants were phenotypically close to other allelic variants of the same gene. Second, through this analysis it was possible to map affinities among pathways due to the phenotypic similarity. Third, the information collected by phenotypic comparison was able to identify orthologous

genes across several species. Again, such results were obtained by the massive comparison of available data, crossing data coming from different sources and formats. Moreover, and interestingly from an epistemological point of view, the order of epistemic steps in the discovery strategy has been inverted. Indeed, while the traditional strategy would have privileged genetic manipulation in order to detect phenotypic variations, here researchers started from the map of known phenotype (a database) and, through a tool capable of integrating such information with other maps, they were able to detect genes and pathways of interest, filtering then possible candidates for further, more classical, experiments. As a matter of fact, such way of doing changes the meaning and the role of experiments themselves in the current practice of science. This fact, being a key aspect of such an epistemological turn, will be precisely addressed in the next section.

In the end, by looking at this kind of science, it should be now obvious how much it embeds a different *way of doing* rather than a change in the theoretical paradigm. Indeed, the molecular tenets are still there. The molecular stance, which drove biomedical research since the 1970s has been certainly modified, definitely extended and revised here and there, but its guiding principle are still valid. This is why I would argue that these new approaches pertain more to the epistemic and methodological side than to the theoretical dimensions of scientific paradigm. They concern how scientific evidences are produced, and how the methods to produce them can be considered reliable and scientific. If traditional molecular biologists were like old fishermen, carefully selecting the bait, the fishing pole, and focused on specific varieties of fish, the new generation of biologists seem to adopt a sort of bottom trawling, trying to collect as much

information as possible. Accidental or not relevant elements such as crabs, prawns, rocks and old shoes (a metaphor for the biological noise) does not constitute a problem, given that the intellectual efforts of scientific practice will shift towards the theoretical principles and the practical constraints of collection design. In the next section I will precisely address the peculiarity of working in science with ontology from an epistemological point view.

Doing science with ontologies: epistemic categories

Molecular biologists usually look for mechanisms (see for instance Craver and Darden 2013). Moreover, as argued in previous chapters, molecular biology lacks a theoretical unification. Molecular biology has then been described more as a set of techniques or, better, experimental cultures (see also Morange 2000, 2006, Rheinberger 1997). Thus I argued that what really makes molecular biology is the adherence of molecular biologists to a certain way of doing. It is a matter of style.

What is then the molecular style? Practice of science is more fluid than theoretical reflection. This is because practices may slightly vary (diversity is a virtue) while theory tends to fill discrepancies. In order to describe a way of doing it is not possible to establish precise necessary and sufficient conditions. Rather, I will try to characterise some notions or hallmarks that clearly circumscribe the practice of molecular biology.

First, experimental systems. The Nobel Prize François Jacob writes that "[i]n analyzing a problem, the biologist is constrained to focus on a fragment of reality, on a piece of the universe which he arbitrarily isolates to define certain of

its parameters. In biology, any study thus begins with the choice of a 'system'" (Jacob 1988, p 234). Experimental systems delimit the purpose, the boundaries and constraints of scientists' research efforts. These systems are constituted by the range of techniques adopted, the types of material instruments and resources, and, of course, the model organism on which the research will be conducted. Experimental systems are then those portions of reality, epistemically and practically demarcated, in which molecular biologists try to make discoveries. However science is not just discovery. Scientists do not just want to number phenomena. They also want to explain them. As already argued in chapter 2, scientific models are the main tool of the explanatory side in science. Thus, in molecular biology, experiments and models are inextricably connected. Indeed, "in molecular biology many experiments serve the purpose of developing and shaping hypotheses – about working models" (Boem and Ratti *forthcoming*). As nicely argued by William Bechtel and Robert Richardson, in order to make the complexity of biological phenomena (that are experimentally addressed) tractable, biologists use models to *decompose* the system into functional or structural elements and then try to *localise* to which structures belong certain functions and vice versa (see Bechtel and Richardson 2010).

The second notion is what Rheinberger (1997) calls *conjuncture*. Accordingly, a conjuncture is the potential intrinsic to the experimental process that can lead scientists to something that was not initially estimated. Following Rheinberger, the discovery of *transfer RNA* is a good example of this aspect. While protein synthesis was originally ad area of pure biochemical investigation, the discovery of such a new molecule made it a central research field in molecular biology. Indeed, the fact that tRNA is a biochemical intermediary

between DNA and proteins, fostered the idea that it could be also an intermediary in genetic information transfer, thus establishing new paths of scientific inquiry.

Third, there is *hybridization*. Such a process occurs when different experimental systems are combined in unforeseen ways. This can reveal unexpected, promising features. "The history of molecular biology is replete with hybridization events. The fusion, *e.g.*, of François Jacob's bacterial conjugation and phage replication system with Jacques Monod's system of induced enzyme synthesis led to the emergence of another novel RNA entity, messenger RNA, and to a pathbreaking model of genetic regulation" (Rheinberger 1997, p s250).

Fourth, *bifurcation*. Briefly, a bifurcation is constituted by a new experimental system stemming out from another one (as when an *in vitro* technique is translated *in vivo*). Sometimes different systems present some degree of sharing, other times they become fully disconnected.

All these elements contribute to create what Rheinberger calls *experimental culture*. As he points out, the adhesion of biologists to such a culture is not determined just by a theoretical commitment (which often is a set of guiding principles imported from other scientific disciplines such as chemistry and physics) but more on material tools and practical behaviours. It is how things are done that best individuates the nature of molecular biology. The seductive metaphor adopted by Rheinberger is that biological research looks then like a net of interconnected experimental systems, deploying different strategies, employing distinct approaches and materials. Namely, the *patchwork view of research*.

I want to argue that the rise of ontologies may challenge this picture. However, more than dismantling it, it is broadening it. A tool like GO has not the purpose to make traditional molecular biology obsolete. It rather changes the meaning that experiments, experimental systems and other categories have for contemporary research. If heuristic strategy of molecular biology is *decomposing* complexity and *localising* its building elements, now ontologies open the possibility to *re-compose* complexity thus adding a new, or at least an additional, layer of what scientific understanding is.

Again, such a change should not be intended as a *paradigm shift*. The point here is to examine what is the peculiarity of doing science with ontologies from an epistemological perspective that takes into account the elements discussed in the previous paragraphs.

First, ontologies seem to extend the notion of experimental system. By the implementation of procedures that allow *packaging* and *un-packaging* data, database seem allowing data to travel (see Leonelli 2009) across different research contexts and experimental systems. In other words, data do not just serve the purposes for which they have been created. They can also be *re-used*. This is certainly true in everyday practice of research. However it is necessary to specify the epistemic nature of such a travel. According to Emanuele Ratti (2015), such a re-use should be intended as a way scientists can pursue in order to establish the presence of common features among different experimental systems. Indeed, following Ratti, data do not simply make a journey across several contexts. The fact that GO provides indications about the type of evidence supporting a given claim, shows that data are not simply packed, unpacked and re-used neglecting their original experimental context. On the

contrary, by creating a *map* that unifies the vocabulary of experimental procedures and resources, ontologies are able to make distinctions across research contexts emerge. Indeed ontologies are enhancing comparison power and not smoothening diversities. This is because they allow data comparison rather than data homogenization. With the use ontologies, the feature of locality of experimental systems is diminished. The "piece of universe" (recalling Jacob's words) isolated by the scientist is not fully confined any longer. On the contrary, it is now always possible to situate the space of experimental manoeuvres into a wider context. In this sense the implementation of ontological work changes also the nature of conjunctures. While in traditional experimental contexts conjunctures have an intrinsic, unforeseen potential for further discoveries which, nevertheless, cannot be disclosed from the beginning, the map provided by a tool like GO makes this epistemic horizon explorable (consider the case of Adamo, Atashpaz, Germain *et al.* 2015) described in the previous section, at least in its directions. Moreover, ontologies modify also hybridization and bifurcation. By standardising the way knowledge is represented, ontologies can either enhance the connections between different experimental contexts or dissolve them. Indeed, the idea of a global map for biological knowledge could mean the end of different epistemic cultures interweaving and contrasting one with another, towards the establishment of a more uniform epistemic scenario. However, again, due to the peculiar form of unification provided by ontologies, I suggest that, rather than suppressing intrinsic and distinctive features of different experimental cultures, ontologies are favouring the appreciation of differences under a common view and not the dissolution of them. As translational dictionaries, ontologies are not conflating different idioms neither reducing one

language into another. They are rather creating a way to grasp the meaning of a sentence (*i.e*. an experimental system) expressed in a given language into another one.

Moreover, *map thinking*, embedded in the application of bio-ontologies, produces a distinctive signature in the way scientific research is thought and perceived, also by scientists themselves. This is because the capacity of ontologies to represent, in a human understandable fashion, the patterns emerging from databases, sets a new frame into which understanding the peculiarity of a prominent part of contemporary research. Indeed ontologies offer a fruitful perspective in order to analyse two important ways of thinking of biological sciences, the *comparative* style and the *exemplary* style, and their epistemic relationship. My claim is that such a distinction is fundamental to understand the peculiarity of many current approaches in doing science.

The two styles embed to different strategies of scientific generalisations of particular findings. While, for instance in comparative anatomy or taxonomy, the generality of a scientific claim is grounded on the *comparison* among many different samples, the discovery of the so-called molecular basis of living things by new biology, promoted the idea that, as famously stated by Monod, "anything found to be true of *E.coli* must also be true of elephants" (Jacob and Monod 1961). This perspective means that, since the 'code of life' has the same structure for all the living beings, the universality of certain finding at the molecular level can be generalised through the assumption that the model organism, taken as the exemplary, serves as a reliable proxy for the phenomenon under investigation. However these two ways of thinking should not be conceived as characterising the disciplinary and epistemic boundaries between natural history and molecular

biology. On the contrary, such a distinction has been proposed by Bruno Strasser and Soraya de Chadarevian (2011) to analyse different components of scientific practices within molecular biology. In their study, Strasser and Chadarevian point out that the historical reconstruction that has depicted the rise of molecular biology as simply the triumph of experimentalism over observations and collection methods employed by natural history, is partially erroneous. Indeed, Strasser and Chadarevian have shown that many great achievements of molecular biology, such as the study of protein structure and function or even the 'crack' of the genetic code, were made possible also because of comparative strategies (think, for instance, about the collections of mutations gathered and classified by Morgan). Molecular biology flourished because of the combination, sometimes even the proficuous contrast, between different styles of reasoning. Very often these styles were anchored to specific phases of the scientific progress. This means that the *exemplary* and the *comparative* style do not represent a way of thinking peculiar to this or that research programme. Rather, these styles were often combined.

The early history of genetics provides a good example of this fact. Indeed, by examining the rise of modern genetics it is possible to detect when and how the generalisation about certain phenomena has been differentially justified by appealing to this or that style. Let us briefly focus on the case of one of the most famous model organism: the fruit-fly *Drosophila melanogaster*. In 1910 Thomas Hunt Morgan "discovered" the first mutant *white eyes* and in 1926 , due to his study on those flies, he published his famous *Theory of the Gene*. Here lies the exemplary style. The theory of Morgan was not only about fruit-flies. The gene became the fundamental unit of biological explanation (see for instance Griffiths

and Stotz 2006, 2013). Every living thing has genes as any material object is composed by atoms. Because of that (and its use in the laboratory work), *Drosophila* has become a symbol of biological research for many experimental biologists. From an experimental point of view, *Drosophila* became really a standard laboratory instrument like a microscope or chemical compounds. However, although fruit-flies were clearly a key component of an experimental work, the way of thinking of Morgan rested also on a very detailed classificatory strategies. Moreover, the capacity of inferring as universals those findings obtained through the fruit-flies was based on the great number of samples and specimens produced and compared. Morgan adopted a first system (called neo-Mendelian) of classifying genetic factors "into organ group systems - eye color, wing shape, body color, thorax pattern" (Kohler 1994, p 56) which helped him to identify "how many genetic factors were involved in the formation of each morphological feature" (*ibid*.). This system was helpful to understand the developmental processes and relationships between different strains. Another classification system Morgan adopted was rather "structural and spatial". The aim of this classificatory approach was useful instead to help scientists to locate physically genetic factors, forming a sort of *genetic map*. Observing, collecting, comparing, were not replaced by the rise of experimental practice, instead they coexisted along with experiments. It is important to notice that not only the practice of classification has been fundamental to complete the genetic study of *Drosophila* but also that different systems of classification provide different answers to questions which often are seen as typically experimental. Moreover, the following failure of the neo-Mendelian system of classification due to the vastness of new mutants, on the one hand forced scientists to elaborate new

classificatory systems and on the other hand helped geneticists to understand the limits of Mendelian genetics. I would say that different ways of knowing have "interfered" with each other. In other words, again, a problem of classification involves directly the practice and the theory of experimental science. However, if we consider the question the other way round, we see how the experimental work affects the strategy of classification. Indeed "drosophilists were the first to encounter the limits of Mendelian system because they were only ones whose breeding experiments *were big enough to produce new mutants*" (Kohler 1994, p 60, emphasis is mine). So choosing a specific tool (*Drosophila*) was the fundamental condition to understand the limits of Mendelian approach. Because of that "Mendelians who worked with mice or fowl had no such experience, because new mutants appeared infrequently if at all in their experiments"(*ibid.*).

Nevertheless, despite this methodological blurriness in which distinct approaches hybridise one into another, the epistemic primacy of experimentalism has definitely prevailed within molecular studies, maybe not entirely in the practice, but certainly in the way the results of biology were publicly disclosed and justified (within and without the scientific community). For instance, an article like the first one examined (Adamo, Atashpaz, Germain *et al.* 2015) would has not been probably published fifteen years ago. This is not because such a study relies on a different theoretical framework, but rather because it employs a diverse working approach. The map thinking shapes the entire rationale of the article allowing to count as evidence what, in the past, was just noise or it could have been considered not relevant. I will come, more in detail, back to this point in the next chapter.

Such an example offers a different perspective according to which interpreting the intellectual battle on the nature of science, mentioned in the first chapter. Indeed, the epistemological point is not just on the adoption of this or that methodology, but rather on the order and hierarchy of distinct ways of thinking. In other words, both opponents in this debate (see, for instance, the controversy on *Nature* 2010 between Robert Weinberg and Todd Golub) do not claim that one scientific practice should entirely replace the other, but they rather state which way of thinking should come first (epistemically, chronologically or economically). Therefore, the rise of ontologies within bioinformatics and their impact on the design of research, should not be understood as a shift from the experimental practice to the advent of a sort of '*in silico* age' of the life sciences. Even if some projects can be certainly pursued purely in a computational fashion, biologists will keep doing experiments. It is not the practice of experimentation that is changing. Rather, it is the transformation of the epistemic role of experiments within research. Thus, such an innovation indicates a difference in the general practice of science. It is a matter of style.

To sum up, in this chapter I provided several examples of current research actually driven by the application of bio-ontologies. I examined different areas of biomedical sciences, by showing how ontologies are not only applied within computational studies, but they also start to be adopted for approaching more traditional problems (such as gene function prediction), offering different and unusual perspectives. Then I proposed an analysis of how bio-ontologies change the practice of science, not just implementing the *comparative* style over the *exemplary* one, but by modifying the hierarchy of methods and evidences of research.

In the next chapter I will examine such a peculiar style of reasoning, compared to traditional molecular biology, and how it affects the epistemic dimension of biomedical sciences.

CHAPTER V

> *"The Scientist must set in order. Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house."*
>
> Henri Poincaré

Molecular biology is dead. Long live to molecular biology

Molecular biology arose in the late first half of the 1900s as an essentially new kind of biology. As already argued and shown, the development of sciences is far from being linear. Following Rheinberger (2007) and Morange (2000) it is possible to identify two main moments or turning point, in the development of molecular biology, which constitute the necessary antecedents in order to understand the cultural and practical change that I address in this study.

In the second chapter I have discussed the problems concerning the status of biology as a discipline. Morange defines molecular biology as "all those techniques and discoveries that make it possible to carry out molecular analyses of the most fundamental biological processes – those involved in the stability, survival, and reproduction of the organisms" (Morange, 2000, p 1). As already argued, it is very hard, if not impossible, to set clear boundaries of such a science, given also that the origin of the molecular paradigm involved the import and the combination of procedures and notions coming from chemistry, physics, genetics (and, in a sense, computer science). Very wisely, rather than numbering or setting necessary and sufficient conditions, Morange individuates, chronologically, some practical and theoretical moves that, produced a sort of

uniformity in the procedural efforts of these new kind of scientists, namely: molecular biologists. Following Morange, in his examination, Rheinberger (2007) adopts the notion of *assemblage* (proposed by the anthropologist Paul Rabinow) to characterise set of elements contributing to the creation and the progress of a scientific discipline as the combination of styles, working strategies and apparatuses, institutions, people and their developmental dynamics. Thus the 'moments' of Morange should be intended, according to Rheinberger, as historical events showing how different components of scientific efforts undergone to a reconfiguration of the *assemblage*.

The first reassembling moment, happened between the 1940s and the 1960s. This phase is mainly characterised by the discovery of the DNA structure and its connection with the study of protein synthesis eventually leading to 'crack' the genetic code (on this aspect, see for instance Darden and Craver 2002). According to Rheinberger, it is possible to individuate some key component for such a change. First, the rise of new technological innovations such as X-ray diffraction, the invention of the electron microscopy, electrophoresis, chromatography and ultracentrifugation. Second, the adoption of model organisms which could be easier manipulated (precisely at the molecular level) than drosophila or maize, such as bacteria (E. coli) and viruses (the bacteriophages). Third, as already stated, the fruitful combination of procedures coming from different disciplinary areas. Fourth, the import, from computer science, of the informational metaphor and its adaptation to biology thus generating the notion of *biological information*.

The second phase happened in the 1970s and its legacy highly affected the 1980s. This phase is heavily characterised by the possibility of actually

manipulate the gene, thus fostering an engineered view of the biological phenomena. Along with *in vitro* analysis, biology went *in vivo*. The development of plasmids, restriction and ligation enzymes and first viral vectors all belong to this period. In the 1980s, the discovery of PCR (polymerase chain reaction) by Kary B. Mullis (see Mullis 1990), which allows the possibility to amplify, virtually, any DNA fragment, extended the range of manipulability in experimental practice. Finally, for molecular biologists, it was possible to intervene on the living cell as a mechanic can work on an engine. Molecular biology as a mature science has been finally established for good.

Key features of molecular biology style: intervention and manipulation

This element can actually help to understand how much manipulation constitutes a key component in the traditional style of reasoning of molecular biology. Certainly in molecular biology the practice of experimentation - the making of science - precedes the theoretical specification at the epistemic level. Following Hacking's idea (1983), experiments have their own life. This is not to say that theory does not play any role in the development of molecular studies, but rather that most epistemological reconstruction that rely too much on theoretical justification fail to entirely grasp the efforts of contemporary biology. Accordingly, it is the manipulation of scientific entities (such as genes) at the experimental level, that grounds the possibility of a more adequate epistemic understanding of what traditional molecular biology is. Indeed, as already pointed out, the process of discovery in molecular biology is far from linear, purely deductive and hierarchical. A key component of the 'molecular style' is then the ability of the single scientist to tinker with the experimental system

allowing him to constantly update and evaluate the relation between the hypothetical assumptions and the empirical results. As Hacking puts it "[i]n schools and colleges experiments are repeated *ad nauseam*. The point of those classroom exercises is never to test or elaborate the theory. The point is to teach people how to become experimenters" (Hacking 1983, p 231). This also shows how much the molecular style privileges practice over theory. It is a way of doing indeed, as the nature of scientific knowledge is conceived, perceived, and conducted as an activity rather than a speculation. Knowing is doing. Therefore *intervening* (i.e. material manipulation) is the main way through which molecular biologists know the world.

As I have shown in the introduction, a style of reasoning sets also the nature of a scientific evidence, its reliability and the validity of the procedures aimed at obtaining it. "Most of molecular biological articles from the mid 1970s are written by telling a story, thereby emphasizing the importance of narratives. The flow of reasoning seems to be as follows. First, one starts from a general guess about a biological system, how it is produced, etc. Due to the generality of the guess, several – though often contrasting – predictions may be derived. Then one devises several experiments exactly to stimulate the experimental system to 'reveal' more information. Some initial predictions are discarded, while others are transformed into hypotheses that are more precise than the initial guess. Next, other experiments are done again to observe reactions from the experimental systems. Then, some hypotheses are further developed while others are discarded. This process continues, virtually, *ad libitum*. This is a sort of progressive and non-linear deductive process, developed by poking and prodding experimental systems. Moreover, in molecular biology, experimental systems are

eminently *created*. Again, as in Hacking's perspective, this is not at all a socio-constructivist drift. In order to study a biological phenomenon, scientists try to isolate it from its environment, creating another – more controlled – context (the so-called experimental system)" (Boem and Ratti, *forthcoming*).

If manipulation is a central feature of the experimental style, it should be notice that experiments may play different roles in the practice of 'questioning nature'. Accordingly, it is possible to distinguish *experiments to prove* (see Popper 1959, Kadane and Seidenfeld 1990) and *experiments to learn* or *exploratory experiments* (see Burian 1997, Steinle 1997). The first ones are those aimed at testing already stated hypotheses while the latter are those capable of fostering the formulation of new hypotheses. However such a distinction (as also argued by Waters 2008) should not be intended in a sharp way. Different types of experiments thus differ not just in their supposed independency from theoretical contributions, but in *how* theoretical constraints affect them. Indeed, different experimental strategies enable different *forms of interventions*. Experiments, whose purpose is to check something that is guided by the theoretical framework, will focus on a narrower range of interventions, thus aiming the attention to those ones which seem most promising according to the understanding of the phenomena under investigation. These experiments are definitely *theory-driven*, in the sense their scope and rationale is constructed according to specific constraints shaped by the theory. However this aspect must be better specified. Following Laura Franklin (2005) here the term "theory" means, at least, two distinct things. On the one hand, theory may count as general *theoretical background*. This should be seen as a broad conceptual stance coupled with empirical findings used as examples. For instance, the original hypotheses

concerning gene regulation and translation, rested on the general idea that protein levels might be controlled by different concentrations of mRNA. On the other hand theory can mean something more specific, narrower and more circumscribed. This is what Franklin calls *local theory*. This term concerns the behaviour of objects which are observed and measured. Consider the Western blot technique. Western blot is a method to determine the presence of a protein of interest in a given sample. In the most common protocol, proteins are first separated according to their molecular weight by running them onto a specific denaturing gel through electrophoresis. Then proteins are transferred from the gel to a nitrocellulose membrane and there the protein of interest is identified by the recognition of a specific antibody. The fact that different bands on the gel are interpreted as showing the distinct molecular weights of different proteins, is part of the *local theory* concerning the Western blot technique. As argued by Franklin (2005), a *theory-driven* experiment is not always directed at the test of a specific hypothesis. Rather, it just requires that such an experiment is designed and performed according to specific theoretical constraints. These constraints are then shaping the possible manipulations, thus narrowing the horizon of expected outcomes. Therefore, the results can meet or not the expectations of the researchers, eventually leading either to confirmations of the previous assumptions, or to methodological problems, or even to potential discoveries. On the contrary, experiments aimed at generating interesting findings about certain phenomena without clearly appealing to a theory, present less formal constraints. In this case the weight of manipulation seems to be higher. Indeed, by employing more intervention possibilities, scientists have also less expectation, because they are 'just' exploring the phenomena. Exploration here means that

experimentation, more than questioning nature, is 'teasing' nature, trying to map how it reacts to material intervention. Of course, such an independence from theory should not be intended in a strong sense, as if scientists did their work completely out of the blue. To give reason to such a 'methodological freedom' we can say that here experiments are only *theory-informed*. Again, the differences in the type of intervention are not in terms of being dependent or not from a theoretical framework, but rather to which degree such a dependence is possible and how it is articulated, i.e. how much and according to which modality, theory affects the practice. Although epistemically useful, such a distinction is not an ontological dichotomy. These divisions should not be intended as fully discrete and mutually exclusive. On the contrary, we should think about different types of experiments as idealised extremities of a methodological spectrum which is quite complex and interrelate. Of course, within a given research programme, it is possible to see different experimental approaches combined.

However, the computational turn, and the production of biological knowledge through the exploration of datasets might let someone to think (as shown before) that now science is facing something different and that the comparative style is actually threatening the old experimental approaches by replacing them. Interestingly, O'Malley (2007) proposed that the increasing role of database and highthrouput techniques in molecular studies should not be seen as mere alternative to experimentation but as a new kind of it, where exemplary and comparative styles are put together. This is why she labels this new approach as *natural history experimentation*. O'Malley examines a very interesting case in which scientists were able to show that a group of marine bacteria can code for a

photoactive protein (*i.e.* proteorhodopsin). The novelty of such a study relies on *metagenomics*. Metagenomics can be defined as the sequencing samples of uncultured micro organisms, taken directly from their environment (see for instance Pace 1997; Chen and Pachter 2005). Since direct hypothesis testing is practically impossible due to the complexity of the samples and their hybrid composition, natural history experimentation "involves various activities of discovery, classification, comparison and probing for specific attributes or properties. Natural history experimentation confers the status of experimenter on nature itself, and reads the results of those experiments as if they had been controlled in biologically meaningful ways […]. More controlled laboratory experiments can, in fact, simply be seen as idealized forms of nature's own experiments. Certain parameters are interpreted as set by nature, and these conditions are taken into account for the systematic comparison of observations" (O'Malley 2007). Following this perspective, if we assume that all these approaches, heavily based on database consultation, data comparison and classification are indeed a new type of exploratory experiments, then it is necessary to clarify in which sense they are different from traditional ones. From a 'style of reasoning' analysis certainly the main aspect regards the notion of intervention as it seems inextricably linked to the practice of experimentation.

If we assume that manipulation, intended as a form of material intervention, is one of the main features of the traditional style in molecular studies, the increasing role of databases in biology and their spread among any level of research, has definitely produced a change in the strategy of molecular science. Although, as already shown in the previous chapter, the practice of comparison has never disappeared from molecular biology, it is also true that the creation of

biological databases, collections of data computationally administered, and thus accessible, have clearly fostered and enhanced the value of comparative approaches as a key mode to generate scientific knowledge. The very term, *data-mining*, suggests this type of turn. Indeed, as miners digging into caves enormous amounts of worthless material in order to find gold and gems, computational scientists penetrate the whole architecture of databases in order to retrieve valuable information. By following Ratti (PhD dissertation, 2015) the "logic of discovery" of these procedures differs from 'traditional' strategies of manipulation substantially. Unlike the search for mechanism, molecular biology here is aimed the individuation of statistically relevant *regularities* within big data sets that only computational approaches can manage. "This access to biological phenomena, through the accumulation of data, has been perceived in a way as 'unbiased' in the sense that phenomena are not created in laboratories by abstracting them from their environment and put into a different context. Instead, data about phenomena are obtained through primary samples. Moreover, there is no need to continuously stimulate experimental systems to squeeze partial information. Though sequencing is a kind of manipulation, data are obtained in one single shot. Moreover, there is no need to devise new experiments to develop general hypotheses, because data obtained by sequencing are taken to be all data one needs in principle. Data are then subjected to bioinformatics analyses, and these analyses do not need additional experiments to put forth hypotheses. Putting aside the initial sequencing part, big consortia do not to need, at first glance, robust interventionist strategies. In other words, sequencing technologies *plus* computational analysis tools, is a kind of molecular biology without the traditional experimental side." (Boem and Ratti, *forthcoming*).

However, I want to argue that the fact the *data-mining* rests on the capacity of exploring vast collections, does not mean that computational methods are not subjected to forms of intervention. On the contrary, the way a collection is constructed heavily affects the type of information that can be retrieved. As argued in the first chapter, a collection is not a mere repository. It is a set of data disposed following a precise order. Ordering and reordering data is definitely a form of intervention, although not material. For the purpose of clarifying this point on collections, let us first consider two examples coming from traditional collections of natural history. This is to show that ordering and reordering is not completely new in the history of the life sciences. Then I will move to contemporary biological databases.

Examples of ordering as intervention

The first example refers to the very origin of modern biology. Talking about the Galapagos' fauna in his *Voyage of the Beagle*, Charles Darwin writes: "It has been mentioned, that the inhabitants can distinguish the tortoises, according to the islands whence they are brought. I was also informed that many of the islands possess trees and plants which do not occur on the others. [...] Unfortunately, *I was not aware of these facts till my collection was nearly completed: it never occurred to me, that the production of islands only few miles apart, and placed under same physical conditions, would be dissimilar*" (Darwin, 1839, 1989, p 287, my italics). Darwin here explicitly says that his collection was almost completed. At that time the final version of his famous theory of evolution was yet to come (the first publication was in 1859). Thus Darwin built his collection

still on a Lamarckian ground. Indeed he put together samples coming from different islands all together. Darwin had to go back to his samples and change the classificatory strategy. He did not change the content, but the general order (we could say the structure of the database) according to which data have been gathered. It is the reordering of that collection according to new theoretical constraints, as nicely described by the historian Giulio Barsanti (2005), that constitutes the first step towards the famous Darwinian theory of evolution.

The second example is more recent. In their famous article *Punctuated Equilibria: An Alternative to Phyletic Gradualism* the two great palaeontologists Niles Eldredge and Stephen Jay Gould (1972) proposed an alternative model to the problem of speciation. While the vast majority of models at that time claimed that evolution of species occurs mainly due to the slow but inexorable and gradual accumulation of modifications, Gould and Eldredge argued in favour of a model of speciation that alternates long periods of stasis with short (in geological terms) cycles of acceleration. It is not relevant here to discuss the details of such a proposal. Rather, the interesting point is that Gould and Eldredge did not construct their model on new experimental findings but they just went back to the, already collected, fossil records. By working with widely known data, but disposing them in a different order, the two scientists were able to highlight new patterns and relations that were invisible to the previous classification.

Although they might seem distant, such intervention procedures are also central in contemporary biological collections, *i.e*. databases. For instance, let us focus on current cancer research and the notion of mutation. *Cancer* is a broad term to refer to a kind of disease characterised by relatively uncontrolled

proliferation of cells that, from a given area of origin, can penetrate into tissues and metastasise to different organs. First examinations under the microscope in the late 19th Century revealed peculiar chromosomal aberrations, thus leading scientists to postulate a central role of hereditary material as a main cause of cancer. The discovery of DNA structure as the molecular ground of biological inheritance eventually promoted the idea that specific agents and environmental contexts can alter the genomic organisation thus generating mutations which can, in turn, give rise to cancer. Despite the variety of tissues in which cancer may occur, the mainstream view (*e.g.* Stratton, Campbell and Futreal 2009) argues that all types of tumour share some common elements. First, the on-going variation at the genetic level in individual cells and second, the selective process on the phenotypic outcomes of such a variation. Speaking roughly, this means that cancer is usually associated with the presence of *mutations*. Here classification plays a central role in mapping knowledge. Indeed, mutations are first distinguished in germline mutations (those inherited by parents) and somatic mutations that are mutations acquired by the cells in their process of differentiations from their progenitors. Nevertheless, further classificatory distinctions can be provided. First one may categorise mutations according to their underlying different biochemical ways. Indeed, given that a genetic mutation is a change in the DNA sequence, this may occur in different manners. First, one nitrogenous base can be substituted by another one. Then, there could be insertions or deletions of DNA segments of different size, or rearrangements, in which DNA breaks in one point and it is subsequently re-joined in another place. Third, the copy number of a gene can exceed the normal number (two, in human diploid genome), as in gene amplification in which the same coding

sequence is repeated several hundred times, or it can be reduced or even deleted from the genome itself. Moreover, the genome can also acquire external genomic content via viral infection (*e.g.* HPV) or even gain epigenetic changes that will affect the structure of the chromatin and consequently the gene expression.

Potentially harming mutations happen regularly in the genome. Most of them however are 'identified' and 'corrected' by the cell machinery. The rate of mutations may vary, depending also on the cell type and on the presence of external factors that can increase their chance of being fixed in the genome itself. Moreover, a global comprehension of the occurrence of mutations is still at an embryonic stage. The picture is also complicated by the fact that cancer progression is not linear and smooth. On the contrary, certain cells can unexpectedly acquire a great number of new mutations with no clear prior indications. As argued in quite recent review "[a]lthough complex and potentially cryptic to decipher, the *catalogue* of somatic mutations present in a cancer cell therefore represents a *cumulative archaeological record of all the mutational processes* the cancer cell has experienced throughout the lifetime of the patient. It provides a rich, and predominantly unmined, source of information for cancer epidemiologists and biologists with which to interrogate the development of individual tumours" (Stratton, Campbell and Futreal 2009, p 720 my italics). It is not by chance that the terms adopted are "catalogue" and "archaeological record". Indeed, despite the obvious differences of data format, the collection of somatic mutations is indeed a type map as the collection of fossils. Indeed, as already argued, collections are *maps* because they highlight the structure, the relational disposition of data, thus creating a common interpretational framework for the information that data convey. The case of

somatic mutations and their classification offers a good example of this aspect. This because here classification does not just organise data, but, by ordering them, it 'creates' phenomena (see again Hacking 1983) as much as material intervention does in traditional experiments. Indeed until 30 years ago there was not any clue that mutations can be separated according to their functional contribution to cancer progressions. In the past, mutations were just different configurations (as described above) of different kinds, concerning genetic sequences. Such a classification is simply structural. When scientists discovered that not all mutations equally contribute to tumorigenesis then the old frame showed its limits. As many other times in its history (see for instance the history of systematics, Mayr 1982, Barsanti 2005) biological world needed to be partitioned in a new way. Moreover, such a new way is not simply the implementation, within the classificatory efforts, of new experimental findings. It should be rather seen the other way round, meaning that empirical puzzles, unsolvable with previous categories, have found a solution through the new order given to data. Let us examine this change.

First, as anticipated, all somatic mutations can be distinguished and categorised according to their effect on tumour development. Usually, those mutations that confer a growth advantage to the cells which possess them are labelled as *driver mutations*. A driver mutation is thought to have a *causal relation* with cancer (*i.e.* cancer is caused by these kind of mutations) and it has been positively selected during the development of the disease itself thus granting its self-sustenance and propagation (although it is not always necessary that a driver mutations is required for the maintenance of tumour till its final stage). Accordingly, driver mutations are found in so called *cancer genes*,

defined as “those genes harboring more mutations than expected, given the average background mutation frequency for the cancer type” (Lawrence *et al.* 2013, p 214). In other words, cancer genes are those genes that, by being suppressed or overexpressed as a result of genomic catastrophic events (*e.g*. mutations or structural variations), can lead to the development of tumours. On the contrary, those mutations that are not involved in cancer progression and maintenance are called *passenger mutations*. Therefore, distinguishing between driver and passenger mutations has become a fundamental task for contemporary cancer genomics.

From a philosophical point of view, such a classificatory approach is particularly interesting as it shows how much the collection design reflects a particular theoretical stance and also reveals some fundamental, underlying, conceptual issues.

First of all, most of the analytical tools to analyse somatic mutations are based on a framework imported by *evolutionary genomics*. This is clear by considering the fact that mutations are divided according to their ability to confer a *selective advantage* to cancer progression. This type of classification can be seen as functional, in the sense commonly ascribed to function in evolutionary studies, that is *selected effect* and that is not immune from serious theoretical debates (see for instance Germain *at al.* 2014). Moreover, the analysis of the rate of mutations can offer a sort of molecular clock to discriminate among different tumour stages, as much as in evolutionary biology it is possible to analyses speciation events (see Vogelstein *et al.* 2013). Accordingly, if cancer cells are the result of driver somatic mutations, and since those mutations confer a growth advantage to cancer cells, then driver mutations are positively selected and they

should be more conserved than mutations having no effect whatsoever on the fitness of cells. Therefore, by increasing the sample size, one would be able to discover those driver mutations that, by conferring growth advantage to cancer cells, are positively selected. However the picture is far more complicated. By analysing the data set of *The Cancer Genome Atlas*, Lawrence and colleagues (Lawrence *et al.* 2013) discovered that such analytical tools (based on the assumption that bigger sample size will lead eventually to the discovery of new cancer genes) needed to be corrected. An analysis of the *whole-exome* (roughly, the coding part) sequence data from 178 lung squamous cell carcinoma revealed that many recurrently mutated genes could be hardly cancer genes. For instance, large genes are notably highly mutated. Moreover, olfactory receptor genes (whose physiological function seems unrelated with lung cancer) are mutated at a suspicious high rate. Thus, scientists decided to see whether taking into account the phenomenon of 'heterogeneity' in tumours can be make sense of such suspicious cases. Heterogeneity in cancer can stand for different things (see Vogelstein *et al.* 2013). Again, classification plays a central role not just in ordering but also in establishing what biological phenomena are as such. First heterogeneity can refer to the fact that mutations, in the same type of tumour, may vary across different patients (*inter-patient heterogeneity*). Second, there is *intra tumoural heterogeneity*, meaning that, within a single primary tumour, cells can show distinct morphological, genetic and phenotypic profiles. Third and fourth, differences can be found between distinct metastases coming from the same primary tumour (*intermetastatic heterogeneity*) and within the same metastasis (*intrametastatic heterogeneity*). Scientists wanted to analyse heterogeneities in 3,083 tumours samples across 27 tumour types of TCGA.

"This analysis does not just advance general knowledge about cancer. Actually, it can also explain the reason why there are certain false positives in using traditional analytical tools for discovering cancer genes. For instance, from the whole analysis researchers revealed that there is a strong correlation between somatic mutation frequency in cancer and low gene expression levels, in the sense that the more a gene is expressed, the less it is mutated. Moreover, they also observe a marked correlation between somatic mutations and DNA replication timing. Two prominent examples of false positive cancer genes analysed by Lawrence and colleagues were olfactory receptor genes. Their high mutation rate is explained by the fact that they are late in replication timing and since they have low expression level in lung tumours, they have a high mutation rate. The same applies to large genes, which in lung cancer are low expressed and are late in replication" (Boem and Ratti, *forthcoming*). Moreover, such a practice of ordering and clustering data in big data sets shows how much classification is a dynamic enterprise. Very often an empirical problem cannot be easily solved precisely because of the categories in which it is framed. Ordering and reordering data in this case show the limits of the cancer gene definition. Besides, classification here can also offer hints on how causal contribution to cancer development can be differently spelled out. Indeed, if cancer genes can be thought as those genes causally linked to cancer onset, such a definition lacks to specify how such a link occurs and whether different ways of contribution might actually exist. In the future, cancer genomics might shift from looking for cancer genes to searching for different and specific *causal factors* in candidate cancer genes. This fact means to reorder data according to new collection principles. First, following Vogelstein and colleagues (2013), it is possible to label cancer

genes as *driver genes* and then to distinguish driver genes from *driver mutations*. Such a clarification is very important since driver genes certainly contain driver mutations but they can also harbour passenger gene mutations. "For example, *APC* is a large driver gene, but only those mutations that truncate the encoded protein within its N-terminal 1600 amino acids are driver gene mutations. Missense mutations throughout the gene, as well as protein-truncating mutations in the C-terminal 1200 amino acids, are passenger gene mutations" (Vogelstein *et al*. 2013, p 1548). However, and more interestingly, Vogelstein and colleagues (*ibid*.) note that those genes which do not contain driver mutations cannot be driver genes by definition. Still, many genes that do not harbour driver mutations can also affect tumorigenesis through their overexpression, underexpression or epigenetic alteration. In a sense, also these genes somehow 'drive the tumour'. In this case a change in the classificatory frame helps to individuate and settle different ways in which a gene can contribute to cancer formation and development. "To reconcile the two connotations of driver genes, we suggest that genes suspected of increasing the selective growth advantage of tumor cells be categorized as either '*Mut-driver genes*' or '*Epi-driver genes*'. Mut-driver genes contain a sufficient number or type of driver gene mutations to unambiguously distinguish them from other genes. Epi-driver genes are expressed aberrantly in tumors but not frequently mutated; they are altered through changes in DNA methylation or chromatin modification that persist as the tumor cell divides" (Vogelstein *et al*. 2013, p 1550).

Conceptual issues

Beside technical problems, a further element of complexity is deeply conceptual. Indeed, all the current approaches heavily rely on the idea that harming mutations mainly occur in the coding region of the genome. This is normally motivated by the fact that a mutation is easily detected when it affects the sequence of an encoded functional product (*e.g.* a protein). However, recent studies (such as Huang *et al*. 2013) have started to show how mutations happening in *intergenic* or *intronic* regions, potentially serving as regulatory elements of the coded part, can also play a central role in tumorigenesis. The *dark matter* (as named by Vogelstein *et al*. 2013) of the genome, *i.e*. the great portion of it that does not contain genes, might press scientists to reconsider their discovery strategies, even promoting a change in the classification that would broaden cancer gene definition itself. Indeed, by changing the notion of the gene, the partition of the genome itself can dramatically change. This can be seen in line with what is argued by ENCODE project authors when they claim that genes should not be seen as fundamental, structurally defined, units of genomic organisations any longer. Rather, according to such a view, "genes represent a higher-order framework around which individual transcripts coalesce, creating a polyfunctional entity that assumes different forms under different cellular states, guided by differential utilization of regulatory DNA" (Stamatoyannopoulos 2012). Assuming a different notion of gene means to also to chose other classification strategies. By changing the category under which biologists classify and order their data, the very same datasets can provide different responses. As for experimental systems, classificatory systems establish the nature of question researchers can ask and thus the kind of answers that can be

obtained. Of course there are some differences. As already argued, molecular biologists can tinker with their experimental system. On this fact rests the possibility of, more or less coherently, setting their space of manoeuvre. Playing with databases is certainly a different game. "Traditional interventionist strategies are clearly different from the practices of 'ordering' in a straightforward sense. Empirical manipulation deals with the 'materiality' of living systems. Even though biological systems are built in the sense that a phenomenon is abstracted from its natural occurrences and 'situated' in a different context, still the experimental systems are subjected to material manipulation. Following Parke (2014), we might say that when a experimenter wants to study one system (the object of study, e.g. lung cancer development in mice) in order to make inferences about another (the target, e.g. lung cancer development in humans) there is the intuition that working on the object puts the experimenter in a privileged position, because there is a sort of 'material' correspondence between the object and the target. Parke (2014) calls this intuition *the materiality thesis*. This thesis has been used to claim in favour of the epistemic privilege of experiments over computer simulations, because in experiments there is a material correspondence between object and target, while in computer simulation the relation is a formal one (see for instance Guala 2002). Because of this materiality thesis, especially in molecular biology, experiments are supposed to have a remarkable inferential power. Ordering data in large datasets clearly does not meet the materiality thesis. As a matter of fact, data are computational 'entities' with certain features having formal relations with each other. Scientists do not materially modify experimental systems when ordering data: they only play with parameters of data to cluster data themselves according

to a certain aim. But this 'playing' is formal, not material. No material manipulations are done in ordering data. However, the materiality of interventionist strategies is just half of the story. 'Intervention' and 'manipulation' imply also that researchers modify a little bit the system under investigation by abstracting some of its features that are of interest for them. Practically speaking, scientists modify some of system's conditions to let certain features emerge. For instance, in the case of disease modelling in murine models, the modification of some genetic features of mice (*e.g.* gene insertion) is aimed at observing phenotypic consequences. 'Ordering' data also implies looking at datasets just with respect to some of its features in order to see what a dataset reveals about itself. However this is not mere 'observation' of data as simply gathered. Actually, database construction means *actively* intervening on the data set by changing some parameters to let emerge only what is of interest. For instance, in the case described above (Lawrence *et al.* 2013), the dataset of TCGA is scrutinised by looking at different types of heterogeneity. Here scientists have 'stimulated' the dataset by clustering data according to a specific aim. Indeed, they have abstracted a dataset from its 'totality' and they have considered just certain features to check the consequences. Dataset considered after ordering is different from the data before ordering. Before ordering, the dataset is just the sum of all data. After ordering, the dataset is what the dataset can tell us about a certain phenomenon. Therefore scientists 'intervene' on the data set, because they want just to observe what is of interest to them by stimulating it to reveal its 'secrets'. To sum up, 'ordering' data (in the sense of clustering) is a form of intervention, though it lacks the 'material' part of typical interventionist strategies of molecular biology." (Boem and Ratti, *forthcoming*).

On the notion of *data*

Nevertheless there is more. Data ordering and reordering is a form of intervention also because data are meaningful precisely because they are disposed and organised according to a particular order. Indeed, data organisation is fundamental not just to comprehend data, but also, and more important, to consider, conceive and perceive them for what they are: *data*. Clarifying this point is crucial. Data are as such only in relation to other data and to the context of their production and gathering. Let us examine how and why.

By following Floridi (2008, 2011) it is possible to distinguish several interpretation of what a *datum* is. First, data can be *epistemically* intended, when they are conceived as collections of facts. This is probably the closest interpretation to the etymological root of the term. *Data* are then 'given' in the sense that they constitute the ground on which constructing further argumentations. Floridi acknowledges that such an account, although useful, lacks in providing an explanations of phenomena as *data compression* and *data cryptography*. Second, data can be equated to *information*. Again, this might be helpful in some practices, but it fails to recognise that the relation is not biconditional, *i.e.* if information "meaningful and truthful data" not every data constitute information. Third, data can be *computationally* conceived as sets of binary elements. This solution however conflates data with the format in which data are encrypted.

In order to overcome all these issues, Floridi adopts what he calls a *diaphoric interpretation* claiming that data stand for, basically, lack of uniformity. A datum is then something that can be recognised, perceived, or

measured as distinctive from the background conditions. However, the relation to the context in which data are produced or gathered is not a simply background. As Floridi writes "[a] white sheet of paper is not just the necessary background condition for the occurrence of a black dot as a datum, it is a constitutive part of the [black-dot-on-white-sheet] datum itself, together with the fundamental relation of inequality that couples it with the dot. Nothing is a datum in itself. Rather, being a datum is an *external property*." (Floridi 2008, p 7 emphasis is mine). If data are relational entities, thus ordering and reordering a database is formally tinkering with the system. It is manipulation. To put it differently, giving order to data is then giving them a meaning. Moreover, each order defines a particular epistemic space. Each data organisation represents a sort of set of classificatory configurations as much as material interventions delimit experimental conditions. "If data are relational entities - in the sense that they acquired their meaning only when they stand in specific relations to each other - then the way we associate a bit of data to another makes the difference as to their interpretation. Contrary to common interpretations portrayed by popular science (see Anderson 2008), data *do not* speak for themselves. In order to give meaning to data, we should organize them in a framework. Mining databases is exactly an operation of putting into specific relations different bits of data. Without ordering, a database is just a sum of data with no meaning. By paraphrasing Hacking, in biological databases there is just complexity, and 'phenomena' (in the sense of meaningful patterns) emerge only if we intervene on the database by reordering it. If in biological databases there is just complexity, after the operation of ordering databases themselves looks quite differently as to the information that we can extract. In this sense, this form of intervention is a kind

of (formal) manipulation because the system (the database) is modified. By relating data in a particular way, we let emerge particular patterns that, strictly speaking, we create by ordering the database. As the Hall effect "does not exist outside of certain kinds of apparatus" (Hacking 1983, p 226), patterns detected through data mining exist only through the algorithm that we apply to mine the database and through the collection designed to build the database. This is a consequence of the notion of *datum* as Floridi meant it, in the sense that data become meaningful only if put in appropriate relations with each other" (Boem and Ratti, *forthcoming*).

Still molecular biology?

The fact that computational approaches and database consultation and curation have changed the practice of molecular biology can mean different things. In a strong sense, one may think that such a 'new' scientific venture has diminished, in some cases even eliminated, the role of experimentation within the research in favour of pure bioinformatics efforts. Although high-sounding, this interpretation is quite inaccurate. As already shown (see the aforementioned case on TGCA or the study on the properties of TS genes presented in Chapter IV), it is certainly true that some projects can be pursued mainly due to the reorder of known data via database integration and consultation. In this sense, mining database constitutes a legitimate form of creating biological knowledge without any contribution from the experimental side. However, it is pretty obvious that the material source of many databases is coming precisely from experimental findings. Certainly, experiments still play a central role in

contemporary research. As anticipated before, it is the role of experiments that has changed. By this, I mean that the contribution of experimental work within the research has epistemically shifted in the very practice of science.

As previously shown, sometimes experiments can be pursued according to their exploratory power. As ancient geographers, molecular biologists were exploring unknown landscapes with no awareness of the configuration of the surrounding areas. It was indeed an exploration. It is not a surprise then that the verb "to explore" has been originally associated with geographical expeditions[22]. However due to the technological advancement, such as satellite technology, also geographical explorations changed their meaning. As a matter of fact, general mapping does not request direct investigation any longer. High-throughput technologies can be seen as the biological counterpart of satellites. However, despite the accuracy of the aerial representations, an investigation of such a kind would inevitably miss some details. These aspects are not fundamental for the global picture but they can become essential for a more complete description of a location. Thus direct expeditions are now aimed not at exploring, but rather at *fitting the details*. Out of this analogy, many experiments in contemporary molecular biology are designed precisely for a similar purpose. This means that experiments, in many cases, do not show an intrinsic aim, but rather they present an instrumental value, as they serve as a confirmation tool. Nowadays, an increasing number of scientific publications perform traditional experiments as confirmations, meaning that empirical results would either corroborate or not the indications provided by high-throughput approaches. For instance, the entire

[22] It is noteworthy to see how *to explore* means: "to go to a country or place in quest of discoveries", thus grounding the activity of discovery on the practice of exploration (see the Online Etymology Dictionary - http://www.etymonline.com/index.php?term=explore)

group of results published by the ENCODE project fits such a picture. As I recently argued with my colleagues (see Germain *et al.* 2014) ENCODE's first step strategy lies precisely in the identification of a "specific subset of biochemical activities (transcription, transcription factor binding, and specific combinations of histone modifications, etc.) which very often contribute and make a difference to the phenomena scientists are interested in" (Germain *et al.* 2014, p 816). At a later stage, these activities will be specified, either confirmed or dismissed via experimentation, which will also clarify their nature and their contribution to the phenomena of interest (by the way, such further efforts do not have necessarily to be conducted by the ENCODE project itself). This means that it is the computational part, not the empirical manipulation, that has assumed the role of exploration, traditionally ascribed to experimentation. In order to better explain this aspect, let us focus on a recent article (Barozzi, Simonatto *et al.* 2014) published on *Molecular Cell* in which traditional experiments are exactly fitting the details, while the general 'questioning nature' is entirely based on computational efforts through database consultation and integration. The aim of the article is to show that *transcription factor binding* is somehow co-regulated with the *nucleosome occupancy* due to the features of certain DNA regulatory segments (enhancers) shared by mammalians. Transcription factors (TFs) are proteins that bind to specific DNA sequences thus regulating the rate of transcription of other functional products. Nucleosomes are instead fundamental units of chromatin organization constituted by a core of proteins, called histones, around which DNA filaments are somehow wrapped up. Enhancers are DNA regions that favour genetic transcription. Nucleosomes are important also for transcription since they contribute to chromatin conformation, thus either

allowing or impeding the possibility of regulatory elements to actually transcribe the genetic information (from DNA to mRNA).

Let us briefly examine the rationale of the study from an epistemic point of view. In this article, scientists start from the knowledge that TFs usually bind sites of regions that previous computational analyses predicted to be with a high nucleosomal occupancy. However, TFs binding sites are hard to detect since their recognition sequence can be easily repeated just by chance, thus creating a high number of false positives. Next, researchers hypothesized that the same information regulating nucleosome establishment also rules TFs in binding specific regulatory elements and neglecting false positives. From an epistemological point of view, all this starting knowledge has been produced by computational approaches. Such approaches were possible because different specific repositories, containing diverse kinds of information (such as factors determining nucleosome occupancy or cell lineage specific enhancers), have been created. Of course this information is as such, precisely because it represents the order according to which data coming from experimental findings have been collected and systematized. Coming back to the article and leaving aside technical details, the important epistemic point here, is that the choices of the experimental system (i.e. what cell lines to work with, what factors to focus on) are fully determined by the needs of bioinformatics. Indeed researchers have chosen to work on primary mouse macrophages and to compare them with several control lines, also because a single specific TF, Pu.1, behaves differently in these diverse cell lineages (it is expressed only in hematopoietic cells). All the experimental materials in this study are instrumental to the accomplishment of the computational analysis or to corroborate *in vivo* and *in vitro* the discoveries

made through bioinformatics tools. Such a change means that experiments here have changed their epistemic role within scientific enterprise. By this I mean that, contrary to the idea that *data-driven* science has diminished the experimental side of scientific work, such a new way of doing rather changes the epistemic primacy of material manipulation that was the benchmark of traditional molecular biology.

By looking at scientific practice in terms of style of reasoning thus, the rise of bioinformatics and biological databases has not enhanced the comparative style, after all never disappeared in molecular biology (see again Bruno Strasser's work, 2012), over the exemplary one. As a matter of fact, the current practice of scientific research in the biomedical field still considers both of them. What has been changed, as I showed, it is rather the relation between the two styles. I would argue that this is due precisely to the epistemic change entailed by *map thinking*. First because maps, by creating a common order for data gathering, provide a general interpretation framework which allows the epistemic journey from data to meaningful information. The second reason involves a higher level of analysis. Indeed, the map thinking also affects the epistemic dimension of science both at its foundational core and at its institutional setting.

Map thinking and molecular biology 3.0

Names and labels have often the purpose to delimitate categories. Thus one may think that 'bioinformatics' well circumscribes the boundaries of a discipline. However, as I argued above many times, when looking at the practice of science we find more styles and approaches deeply intertwined one into another, rather than clean and sharp disciplinary limits. From this perspective, disciplines

themselves look like *a posteriori*, epistemic reconstructions. Thus the change promoted by the rise of computational approaches and the creation of databases does not constitute a shift from molecular biology to a new discipline but rather an epistemic reordering of styles within the biomedical research field which, as shown in the previous chapters, is still grounded on the theoretical paradigm at the very origin of molecular biology. However, something else changed too.

By adopting the evocative notion of *assemblage*, as a sign of a non hetero directed phenomenon, Rheinberger suggests that "the coming into being of molecular biology was certainly not a *project* in the sense of, for instance, the *humane genome initiative* in the late 1980s" (Rheinberger, 2007, p 218, emphasis is mine). Certainly, the rise and the development of molecular science was possible also (someone could say mainly) due to economical and political decisions. Indeed, the central role of the *Rockefeller Foundation* in shaping the form and the aims of biological research at the beginning of the molecular era, is not a mystery to anyone (see Morange 1998, 2000; Strasser 2014). Also the making of *biotech industry*, which corresponds to the aforementioned second phase, presents a strong political and societal drive. The historian of science and technology Eric Vettel has recently reconstructed the development of biotechnology by linking it to the cultural and political scenario of the 1970s in the United States (Vettel, 2006). Vettel analysis sheds a light on two important aspect of the development of molecular studies after the first phase. First, while early molecular biology rested mainly on philanthropic funding (*i.e.* the Rockefeller Foundation) and on pioneering efforts of scientists coming from different research fields, the second shift has been instead characterised by the creation of specific federal funding, thus also helping to delineate and stabilise

molecular biology as a discipline. Indeed, while many research groups working with molecules were usually part of medical schools or agricultural departments, such a shift let some peculiar laboratories to emerge as genuinely independent because of their specific research agenda. In particular, the Wendell Stanley's Biochemistry and Virus Laboratory in Berkley, arose as a messenger of a new kind of science focused on basic research, and because of that connected with the epistemic idea (and rhetoric) of pursuing 'pure knowledge', which highlighted the link with already established 'noble' sciences, such as physics. This aspect is important to understand that the institutional independence of Stanley's lab allowed also epistemic independence meaning that the procedures, the methodologies and styles employed by the lab were not subjected to higher authority. However, as already argued in previous chapters, science development is far from being linear and straightforward. Paradoxically, it was Stanley's lab emphasis on its work concerning the physics and the chemistry of life that granted the possibility of one of the most technological innovations in the life sciences: *recombinant DNA*. Such a technique allows the creation of specific molecules, in order to combine together genetic material from diverse sources, thus generating sequences that would not otherwise be found in biological organisms. The very possibility of recombinant DNA lies in the fact that all known living beings share the same DNA structure. Indeed, a message perfectly in line with Stanley Laboratory's views. However, if in some places some research directions are settled, the very same decisions can also promote *countercultures*. As argued by Doogab Yi (2015), the advent of recombinant DNA technique fostered the rise of biotech companies which creates, concomitantly, a new kind of researchers, both scientists and entrepreneurs. Far

from being clean and clear, this period was full of tensions, both at the social and the epistemic level. This was particularly true in the relationship between the life sciences and medicine. "Molecular biology was challenged at the institutional level by those who opposed its 'imperialistic' disciplinary politics and its standing as a new postwar discipline. […] Often, molecular biologists' scientific and medical claims were bold enough to draw criticism from researchers in a number of other biological fields. To some molecular biologists, it seemed that the future of their discipline increasingly depended on its ability to find available intellectual and institutional niches in order to become a constructive part of the expanding biomedical complex" (Yi 2015, p 61-62). In this sense, in the same years, the US government promoted policies emphasising the practical implementation of biological findings. Such a passage helps to understand the transition from the life sciences research to biomedicine. Accordingly, the famous *War on Cancer* supported by Richard Nixon administration (National Cancer Act of 1971) in those crucial years, has definitely contributed to create the field of biomedical research. As also argued by Strasser "[a]t present, 'biomedical research' is used […] to designate a form of medical research based on experimentation in the laboratory and framed by knowledge in natural sciences, such as physiology or bacteriology" (Strasser 2014, p 11). In terms of style of reasoning this means that it is the practice of experimentation, the experimental way of doing, meaning the conviction that material manipulation allows the best understanding of phenomena, that has framed the change from early molecular biology to the second phase that Rheinberger (2007) calls the "gene technological shift". Again, the molecular paradigm, the view that living phenomena can be rightfully grasped and described at their best at the molecular

level still holds.

The new turning point, as also recognised by Rheinberger, is the Human Genome Project. The HGP constitutes (see chapter 2) a first rupture or the beginning of a new phase. This is true in many senses. First, as already argued in the second chapter, because the HGP reflects and re-establishes a way of thinking based on the creation of maps and the fact that the knowledge produced through maps would have solved some key problems, that were considered un-attachable from other approaches. In other words mapping would reveal secrets that other methods will inevitably miss. "The voices that placed genome sequencing on a par with a march to the holy grail of life became loud and dominant" (Rheinberger 2007, p 221). Second, because the shift towards map building coincides with the establishment of big science projects. Contrary to the *assemblage*, with the HGP, biomedical research seems to have found a higher order which determines its methods, approaches and epistemic desiderata. The proliferation of big science projects over the dispersed frame of traditional laboratories, is a challenge to the organisation of scientific venture also from a genuine epistemological point of view. Indeed, this situation challenges also philosophy of science itself. As a matter of fact, such a transformation confronts both those who have defended the irreducibility of scientific discovery to its rational, *a posteriori*, systematisation (see, among all, Feyerabend 1975), and those who have advocated for intrinsic value of epistemic pluralism as the main sign of the success of natural sciences (*e.g.* Dupré 1993). In *Against Method* Feyerabend argued that any attempt to individuate *a* method of scientific enterprise is doomed to fail. In other words, his famous *anarchist* stance claims that there are no suitable, exception-less, methodological procedures regulating

the development of science or the advancement of knowledge. Moreover, any effort to establish or impose a set of features designating the method of science will produce the undesired effect of inhibiting scientific progress itself, by applying too narrow and exclusive conditions. On the other side, Dupré (1993) challenges the idea that knowledge can be equated to the attempt to provide a consistent, general order to the chaos of everyday experience. In doing so, he endorses a form of epistemic pluralism concerning scientific method, but also, and more profoundly, he defends the idea that the very term 'science', conceived as a unique endeavour, reveals a metaphysical assumption which can be detrimental for the message of scientific disciplines. Moreover, as recently argued by Kyle Stanford, "Kuhn himself also argued influentially that even in the course of such normal science the intellectual flexibility and freedom of younger scholars and those new to a given scientific field to propose and pursue alternatives to existing theoretical orthodoxy was the most crucial ingredient in the possibility of any truly fundamental or revolutionary change in our scientific beliefs" (Stanford 2015, p 9). All these arguments defend the freedom of science and its intrinsic pluralism as the key of its success.

The establishment of *map thinking* seems to contest the epistemic freedom of science at different levels.

First, this involves the nature of scientific explanations. The change from the *gene* to the *genome* (and next, all the other '-*omes*') it is not only a mutation in the object of scientific inquiry that obviously requires the adoption of new tools and discovery strategies, but also, as already shown, a transformation in the style of reasoning. This means that what counts as scientific data and reliable evidence will be prone to the peculiar explanatory strategy of big science projects. By this

I mean that the independence of local explanations, usually those provided by mechanistic molecular biology, will be inevitably diminished, because all empirical results would be as such only in the wider, common context provided my maps of biological knowledge. In the age of map thinking only the general picture, no matter if still inaccurate or incomplete, could provide the most genuine level of explanation. Indeed, by broadening the notion of experimental systems, and contrary to what argued by Hacking (1983), maps seem pointing at phenomena not by revealing them in *ceteris paribus* conditions, but by putting different sources in relation for the construction of a common picture. The take home message is clear: true scientific explanations are *global explanations*. Paraphrasing Dobzhansky's dictum (Dobzhansky 1964) one may argue that "nothing in molecular biology will make sense except in the light of map building". Traditional molecular biology becomes then explanatory flawed because of its reductionist stance. Indeed "[i]n the late 1990s, scientists, among them Ludwig Winnacker, who was president of the German Research Foundation (DFG) at the time, talked of the beginning of a new age, the age of postgenomics. Statements began to be heard such as 'It is time to transcend old reductionisms,' and 'What must come into view again is the whole of the organism in the full broadness of its functions on the level of cells, tissues, and organs and in the depth of its development'" (Rheinberger 2007, p 221). Of course, in Winnacker's words there is a certain amount of rhetoric. As previously argued, mapping does not solve, *per se*, the problem of reduction. The 'global' perspective granted by highthrouput approaches cannot be equated to the holistic understanding of the underlying phenomena. Indeed, roughly speaking, holism stands for the idea that the whole cannot fully correspond to the mere sum of its

parts. I am not interested in entering such a debate, but it should be clear that *map thinking*, as shown in chapter 3, has certainly the capacity to look at different parts in an integrated manner (due to the shared representational framework it provides) but cannot resolve complexity by itself (see also Morange 2006). Maps, as previously argued, constitute a form of unification that should be labelled as *standardisation*. Such a standardisation is the key feature according to which data coming from different sources can be compared and integrated. This also explains why maps 'work' even if they are not complete or fully accurate. The most important feature of maps is to highlight the underlying structure which connects different kinds of data. However, standardisation here means also that data should be produced the more and more according to shared procedural rules and common *sanity check parameters* (see Stevens 2013).

Second, from the institutional side this has an obvious consequence. As nicely shown by the biologists Aaron Hirsch "[a]cross many different fields, new data are generated by a smaller and smaller number of bigger and bigger projects.[…] If the nineteenth century was an age of far-flung investigators alone in the wilderness or the book-lined study, the twenty first century is, so far, an age of scientists as administrators. Many of the best-known scientists of our day are men and women exceptionally talented in herding the resources—human and otherwise—required to plan, construct, and use big sophisticated facilities.[…] There's something disturbingly *hierarchical about the new architecture of the scientific community*: what was before something like a network of small villages is today more like an urban high-rise, with big offices at the top and a lot of cubicles down below" (Hirsch 2009, emphasis is mine). The epistemic consequence of this can surprisingly offer and pragmatic and sharp way to

establish what science is from what is not, or at least to separate what I can label as *major league science* - the 'good one' – from minor and marginal scientific efforts. Under the umbrella of such a methodological homogenisation, the disciplinary boundaries within molecular studies will vanish as the new biology will take place. If the rise of bioinformatics – and its impact on the practice of research - can be labelled as biology 2.0, the epistemic primacy of classificatory strategies and of the comparative style through the construction of immense databases, over the traditional material manipulation of wet biology experiments can be seen as biology 3.0. As Hallam Stevens recently argued "[w]e can already perceive the outlines of what this biology might look like. First, it draws on the tools of Web 3.0, particularly the Semantic Web, to create a hyperdata-driven biology. Not only will massive amounts of biological data be available online (this is already true), but these data may soon be semantically connected in such a way that discoveries about biological function can readily fall out of the data. But Biology 3.0 also constitutes an erasure of the boundary between the biological and the informatic: biological objects and their informatic representations will become apparently interchangeable. Data will be rich and reliable enough that doing a digital experiment by manipulating data will be considered the same thing as doing an experiment with cells and molecules.[…] Biology 3.0 predicts the culmination of the process in which the biological and the informatic have become a single practice. In other words, bioinformatics may disappear. The practice of using computers to generate or run simulations in physics is not designated separately from the rest of physics—there is no "phys-informatics." Such could be the case for bioinformatics—its practices seem likely to become so ubiquitous that it will be absorbed into biology itself. More

precisely, what the notion of Biology 3.0 suggests is that the practices and knowledge associated with bioinformatics may gradually subsume those of the rest of biology.[…] The 'wet' work of biology may become increasingly confined to highly ordered and disciplined spaces designed to produce data with the greatest possible efficiency. Meanwhile, "dry" biology can be done anywhere, by anyone with a computer and an Internet connection. (Stevens 2013, p 219-220).

If this future depicted by Stevens will happen, it will be precisely because of the role of biomedical ontologies. As tools which overarch the boundaries of specific databases, ontologies are the first and the best candidate to manage the information coming from different sources and to integrate them - through their standardisation power – thus creating the most comprehensive and updated map of current biological knowledge. However ontologies can do even more.

As previously shown in chapter 3, ontologies were originally intended as tools of integration. Yet, recently they have started to be employed for *scientific*, *technological and medical publishing* (STM) in order to increase information gathering and text mining thus enhancing application for hypotheses generation and promoting common assets for discovery strategies. The application of ontologies to scientific literature in order to create a standard in information retrieval has been recommended by various scholars (see Blake 2004, Seringhaus and Gerstein 2007) in the biomedical field. An example is offered by the tool Textpresso (Müller, Kenny, and Sternberg 2004) which is an ontology-based mining and retrieval system. Once a set of articles has been uploaded, Textpresso is able to identify single sentences and to relate them to 33 ontological terms, three of which are GO ontologies. The results are then ranked according to their

relevance by highlighting the terms and can be visualised in a single chart that provides also links to external databases. Even if scientists will still continue to read articles in a traditional manner, these tools will increasingly shape the way they gather technical information by creating an always more refined literature which will be fully integrated in the ontological paradigm. Moreover "formalized assertions, perhaps maintained in specialized "structured abstracts" will provide indexing and browsing tools with computational access to causal and ontological relationships. Hypertext linking will be extensive, generated both automatically and by readers providing commentary on blogs and through shared annotation databases. At the same time, more tools for enhanced searching, scanning, and analyzing will appear and exploit the increasingly rich layer of indexing, linking, and annotation information" (Reaner and Palmer 2009, p 832). However fascinating, and oriented to increase the efficiency of information retrieval, this scenario has also a potential dark side. Analogously with Amazon or Google algorithms, which prioritise both research results and the type of information according to users' profile, such a unification of research strategies can potentially reduce the freedom of single scientists (who can also be totally unaware of this).

CONCLUSIONS

In a very recent *Nature*'s book review of the historian David Wootton's last manuscript (*The Invention of Science: A New History of the Scientific Revolution*, Penguin, UK, 2015) it is written that "[w]hat marks out modern science is not the conduct of experiments, […] but the formation of a critical community capable of assessing discoveries and replicating results" (Ball 2015, p 413). Making experiments has been thought and perceived indeed as one the hallmarks of modern science, and molecular biologists considered their discipline epistemically mature and genuinely scientific (such as physics) precisely because of the power of experiments. However, it is not the material manipulation in itself, that has granted the real success to molecular biology. It was rather a matter of style. In other words, the main feature of molecular biology was the establishment of a community which adopted *experimentalism* (*i.e.* the idea that 'truths about nature' can be discovered and justified through specific tests conducted under particular and controlled conditions) as a way of working and thus, as a way of thinking. Indeed, such a way guided biologists in determining what counts as evidence, explanation or what constitutes a reliable solution to a problem.

Hence, *styles of reasoning* are a powerful analytic tool to examine, dissect and track the changes in the development of a scientific discipline. The fact that in molecular biology the role of computers has definitely grown, cannot be seen, *per se*, as the element deciphering the essence of the transformation through which computational methodologies are affecting molecular research. The rise of bioinformatics reveals more. It shows that change of practice is not simply a methodological and technological innovation but, more important, it is rather a

change in the way scientists decide what a good scientific practice is. Moreover, it is also a change in the meaning of key scientific notion such as *proof*, *evidence*, *explanation* and *experiment*. It is a transformation in how molecular biologists think.

In this study, I tried to show that *map thinking* precisely embraces and resumes at best the features of this new way of knowing. This is because the notion of map embeds some features of the classificatory styles under a new light. In doing so I tried also to highlight how much of this novelty actually rests on styles and epistemic cultures already present in the history of sciences but filled with totally new elements. Indeed, contemporary map thinking in molecular biology should not be seen neither as the mere presenting again of old approaches nor as something which determines the disappear of experimentalism from molecular studies. This because styles of reason are epistemic categories which have to be intertwined with other analytic tools, such as *models*, *metaphors* and *paradigms*. Therefore, rather than simply opposing the novelties provided by new computational approaches I tried to show how map thinking has restructured, rather than dissolved, the role of traditional elements of research according to new and different epistemic criteria.

The value of this work, I think, lies also in the fact such an epistemological reconstruction is not simply based on philosophical analysis but is also grounded on the rise of specific tools, *bio-ontologies*, which definitely incarnate the spirit of map thinking. The intrinsic encyclopaedic agenda of bio-ontologies perfectly represents the empirical counterpart of my theoretical discourse. This explains why I dedicated to bio-ontologies a large part of my attention. Ontologies are not just a tool. They are really a different way of knowing in practice with the

potential to change the face of contemporary molecular research. Would this change be positive for science? Despite its intrinsic interest, a discussion on values within science is beyond the scope of this thesis. Moreover, the meaning of "positive" here is not so easy to address. However, from a philosophical perspective, some genuine epistemic points can be stressed out.

As I already argued elsewhere, "[s]uch a change has the potential of being more disruptive, at the epistemic level, than one may think. Indeed, this new venture seems to challenge the idea, supported by the second phase of philosophy of science, according to which there is no logic of discovery within sciences, since scientific enterprise cannot be completely reduced to clean and sharp logical steps (Feyerabend 1975). The richness of science has been argued as lying precisely in its capacity of going through different paths. Thus epistemic pluralism (Dupré 1993) has been established as the current mainstream view concerning the success of science and also about the intrinsic value of scientific research itself. However, the constraints imposed by consortia running Big Science projects do not pertain just the economic side of scientific research. By creating a standardization of methods and procedures, Big Science projects are, inevitably and probably unwittingly creating, from the practical side, a set of criteria for the old *demarcation problem*. If the technical and the epistemic repertoire of molecular studies will be completely subdued to the creation of vast maps of biological knowledge through computational approaches of massive amounts of data, scientific methodological pluralism will pass over. *Pace* Dupré's "disorder of things," it seems that a new order could prevail. If such a thing will happen, at the moment is just a risk. This risk is twofold. On the one hand no one knows whether such a standardization will contribute either to shape

a better science (in the sense of more effective) or to impoverish it[23]. On the other hand, if such a way of doing should take place more vastly in the research landscape, this would also mean the end of many small labs that will not be able to do science any longer. Perhaps the core of the problem rests on the notion of 'effectiveness'. What does it mean to be effective in science? How is this related with the idea that scientists and philosophers have about what "good science" is? Such a concept because of its importance for scientific research, should be one of the future most important challenge for both science and philosophy of science" (Boem and Ratti, *forthcoming*)

[23] For instance, standardisation could undermine the creative aspects of scientific discovery which, quite paradoxically, are sometimes taken as important as the rational nature of the scientific work

BIBLIOGRAPHY

Adamo, A., Atashpaz, S., Germain, P.-L., Zanella, M., D'Agostino, G., Albertin, V., … Testa, G. (2015). 7q11.23 dosage-dependent dysregulation in human pluripotent stem cells affects transcriptional programs in disease-relevant lineages. *Nature Genetics*, *47*(2), 132–41. http://doi.org/10.1038/ng.3169

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Retrieved August 29, 2015, from http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/

Aristotle (2012). *The Metaphysics*. Roger Bishop Jones.

Aristotle (2000). *Nicomachean Ethics (Second Edition)*. Hackett Publishing

Ashburner, M., Ball, C. a, Blake, J. a, Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. http://doi.org/10.1038/75556

Ball, P. (2015). History of science: The crucible of change. *Nature*, *524*(7566), 412–413. http://doi.org/10.1038/524412a

Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., & Natoli, G. (2014). Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Molecular Cell*, *54*(5), 844–57. http://doi.org/10.1016/j.molcel.2014.04.006

Barsanti, G. (1992). *La scala, la mappa, l'albero: immagini e classificazioni della natura fra Sei e Ottocento*. Sansoni.

Barsanti, G. (2005). *Una lunga pazienza cieca: storia dell'evoluzionismo*. Einaudi.

Bechtel, W., & Richardson, R. C. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. MIT Press.

Bertolaso, M. (2015). *The Future of Scientific Practice: "Bio-Techno-Logos."* Taylor & Francis.

Bird, A. (2007). *Nature's Metaphysics: Laws and Properties*. Clarendon Press.

Blake, J. (2004). Bio-ontologies-fast and furious. *Nature Biotechnology*, *22*(6), 773–4. http://doi.org/10.1038/nbt0604-773

Boem, F., Pavelka, Z., & Boniolo, G. (2015). Stratification and Biomedicine: How Philosophy stems from Medicine and Biotechnology. In M. Bertolaso (Ed.), *The Future of Scientific Practice: "Bio-Techno-Logos."* Pickering & Chatto.

Boem, F., & Ratti, E. (n.d.). Ordering and re-ordering data: Towards a Notion of Intervention in Big-Data Science. In M. Nathan & G. Boniolo (Eds.), *Philosophy of Molecular Medicine*. Routledge.

Boniolo, G. (2007). *On Scientific Representation: From Kant to a New Philosophy of Science*. Palgrave Macmillan.

Boniolo, G. (2013). On Molecular Mechanisms and Contexts of Physical Explanation. *Biological Theory*, *7*(3), 256–265. http://doi.org/10.1007/s13752-012-0073-z

Brenner, S. (2002). Life sentences: Ontology recapitulates philology. *Genome Biology*, *3*(4), comment1006.1–comment1006.2. http://doi.org/10.1186/gb-2002-3-4-comment1006

Brenner, S. (2010). Sequences and consequences. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*(1537), 207–12. http://doi.org/10.1098/rstb.2009.0221

Burian, R. M. (1997). Exploratory Experimentation and the Role of Histochemical Techniques in the Work of Jean Brachet, 1938-1952. *History and Philosophy of the Life Sciences*, *19*(1), 27–45.

Burian, R. M. (2007). On MicroRNA and the Need for Exploratory Experimentation in Post-Genomic Molecular Biology. *History and Philosophy of the Life Sciences*, *29*(3), 285–311.

Callebaut, W. (2012a). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 69–80. http://doi.org/10.1016/j.shpsc.2011.10.007

Callebaut, W. (2013). Naturalizing Theorizing: Beyond a Theory of Biological Theories. *Biological Theory*, *7*(4), 413–429. http://doi.org/10.1007/s13752-013-0122-2

Callinan, P. a., & Feinberg, A. P. (2006). The emerging science of epigenomics. *Human Molecular Genetics*, *15 Spec No*(1), 95–101. http://doi.org/10.1093/hmg/ddl095

Cambrosio, A., Keating, P., Schlich, T., & Weisz, G. (2006). Regulatory objectivity and the generation and management of evidence in medicine. *Social Science & Medicine (1982)*, *63*(1), 189–99. http://doi.org/10.1016/j.socscimed.2005.12.007

Carey, N. (2012). *The Epigenetics Revolution: How Modern Biology Is Rewriting Our Understanding of Genetics, Disease, and Inheritance*. Columbia University Press.

Carroll, J. W. (1994). *Laws of Nature*. Cambridge University Press.

Carroll, S. B. (2006). *Endless Forms Most Beautiful: The New Science of Evo Devo*. W. W. Norton.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Clarendon Press.

Cassirer, E. (1910). *Substance and Function*. Dover Publications.

Chadarevian, S. de. (2009). Interview with Sydney Brenner. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *40*(1), 65–71. http://doi.org/10.1016/j.shpsc.2008.12.008

Chaffer, C. L., Marjanovic, N. D., Lee, T., Bell, G., Kleer, C. G., Reinhardt, F., … Weinberg, R. A. (2013). Poised chromatin at the ZEB1 promoter enables breast cancer cell plasticity and enhances tumorigenicity. *Cell*, *154*(1), 61–74. http://doi.org/10.1016/j.cell.2013.06.005

Chen, K., & Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Computational Biology*, *1*(2), 106–12. http://doi.org/10.1371/journal.pcbi.0010024

Cheng, L., Lin, H., Hu, Y., Wang, J., & Yang, Z. (2014). Gene function prediction based on the Gene Ontology hierarchical structure. *PloS One*, *9*(9), e107187. http://doi.org/10.1371/journal.pone.0107187

Christofferson, D. E., & Yuan, J. (2010). Necroptosis as an alternative form of programmed cell death. *Current Opinion in Cell Biology*, *22*(2), 263–8. http://doi.org/10.1016/j.ceb.2009.12.003

Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. University of Chicago Press.

Crombie, A. C. (1994). *Styles of Scientific Thinking in the European Tradition: The History of Argument and Explanation Especially in the Mathematical and Biomedical Sciences and Arts, Volume 2*.

Darden, L., & Craver, C. (2002). Strategies in the Interfiled Discovery of the Mechanism of Protein Synthesis. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *33*, 1–28.

Darden, L., & Craver, C. (2002). Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *33*(1), 1–28. http://doi.org/10.1016/S1369-8486(01)00021-8

Darity, W. A. (Ed.). (2008). *International encyclopedia of the social sciences* Macmillan Reference USA/Thomson Gale.

Darwin, C. (2015). *The Voyage of the Beagle*. Sheba Blake Publishing.

Darwin, C., & Keynes, R. D. (2001). *Charles Darwin's Beagle Diary*. Cambridge University Press.

Daston, L., & Galison, P. (1992). The Image of Objectivity. *Representations*, *40*(40), 81–128. http://doi.org/10.2307/2928741

Daston, L., & Galison, P. (2010). *Objectivity*. Zone Books.

Davis, B. D. (1992) Sequencing the human genome: a faded goal. *Bulletin of the New York Academy of Medicine*, *68*(1), 115–25.

Davis, B.D. (1991). *The Genetic revolution: scientific prospects and public perceptions*. Johns Hopkins University Press.

Dennett, D. C. (1995). *Darwin's Dangerous Idea: Evolution and the Meaning of Life*. Simon & Schuster.

Diamond, J. (1999). *Guns, Germs, and Steel: The Fates of Human Societies*. W. W. Norton.

Dobzhansky, T. (1964). Biology, molecular and organismic. *Integrative and Comparative Biology*, *4*(4), 443–452. http://doi.org/10.1093/icb/4.4.443

Dröscher, A. (2008). *Biologia. Storia e concetti*. Carocci.

Dupré, J. (1983). The Disunity of Science. *Mind*, *XCII*(367), 321–346.

Dupré, J. (1993). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Harvard University Press.

Dupré, J. (2001). In defence of classification. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *32*(2), 203–219. http://doi.org/10.1016/S1369-8486(01)00003-6

Eldredge, N., & Gould, S. J. (1972). Punctuated Equilibria: An Alternative To Phyletic Gradualism. *Models In Paleobiology*. http://doi.org/10.1037/h0022328

Feyerabend, P. (1993). *Against Method*. Verso.

Floridi, L. (2008a). Data. In W. Darity (Ed.), *Encyclopedia of the Social Sciences*. Macmillan Publishers Limited.

Floridi, L. (Ed.). (2008b). *The Blackwell Guide to the Philosophy of Computing and Information*. John Wiley & Sons.

Floridi, L. (2011). *The Philosophy of Information*. OUP Oxford.

Foster, M. (1899). *The Growth of Science in the Nineteenth Century*. British Association for the Advancement of Science.

Fraassen, B. C. Van. (1980). *The Scientific Image*. Clarendon Press.

Fraassen, B. C. Van. (2008). *Scientific Representation: Paradoxes of Perspective*. OUP Oxford.

Franklin, L. R. (2005b). Exploratory Experiments. *Philosophy of Science*, *72*(5), 888–899. http://doi.org/10.1086/508117

Germain, P. L., Ratti, E., & Boem, F. (2014). Junk or functional DNA? ENCODE and the function controversy. *Biology & Philosophy*.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., … Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, *17*(6), 669–81. http://doi.org/10.1101/gr.6339607

Giardino, V. (2013). Towards a diagrammatic classification. *The Knowledge Engineering Review*, *28*(03), 237–248. http://doi.org/10.1017/S0269888913000222

Giere, R. N. (1999). *Science Without Laws*. University of Chicago Press.

Giere, R. N. (2006). Perspectival Pluralism. In S. H. Kellert, H. E. Longino, & K. C. Waters (Eds.), *Minnesota Studies in the Philosophy of Science* (pp. 26–41). University of Minnesota Press.

Giere, R. N. (2010). *Scientific Perspectivism*. University of Chicago Press.

Gilbert, W. (1991). Towards a paradigm shift in biology. *Nature*. http://doi.org/10.1038/349099a0

Godfrey-Smith, P. (2013). *Philosophy of Biology*. Princeton University Press.

Golub, T. (2010). Counterpoint: Data first. *Nature*, *464*(7289), 679. http://doi.org/10.1038/464679a

Gould, S. J. (1996). *Full House*. Harvard University Press.

Gould, S. J. (2002). *The Structure of Evolutionary Theory*. Harvard University Press.

Gould, S. J. (2009). *Punctuated Equilibrium*. Harvard University Press.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., & Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution*, *5*(3), 578–90. http://doi.org/10.1093/gbe/evt028

Grenon, P., Smith, B., & Goldberg, L. (2004). Biodynamic ontology: applying BFO in the biomedical domain. *Studies in Health Technology and Informatics*, *102*, 20–38.

Griffiths, P. E. (2008). *Ethology, Sociobiology, and Evolutionary Psychology*. *A Companion to the Philosophy of Biology*. http://doi.org/10.1002/9780470696590.ch21

Griffiths, P. E., & Stotz, K. (2006). Genes in the postgenomic era. *Theoretical Medicine and Bioethics*, *27*(6), 499–521. http://doi.org/10.1007/s11017-006-9020-y

Griffiths, P., & Stotz, K. (2013). *Genetics and Philosophy: An Introduction*. Cambridge University Press.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. http://doi.org/10.1006/knac.1993.1008

Gruber, T. R. (2009). What is an ontology? In L. Liu & T. Özsu (Eds.), *Encyclopedia of Database Systems*. Springer-Verlag.

Guala, F. (2002). Models, Simulations, and Experiments. In *Model-based reasoning*. *Science, technology, values* (pp. 59–74). http://doi.org/10.1007/978-1-4615-0605-8_4

Gurdon, J. B. (1962). The Developmental Capacity of Nuclei taken from Intestinal Epithelium Cells of Feeding Tadpoles. *J Embryol Exp Morphol*, *10*(4), 622–640.

Hacking, I. (1983). *Representing and Intervening*. Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9780511814563

Hacking, I. (1985). Styles of Scientific Reasoning. In J. Rajchman & C. West (Eds.), *Postanalytic Philosophy*.

Hacking, I. (1994). Styles of scientific thinking or reasoning: A new analytical tool for historians and philosophers of the sciences. In K. Gavroglu, J. Christianidis, & E. Nicolaidis (Eds.), *Trends in the historiography of science*. Kluwer Academic Publishers.

Hacking, I. (2002). Language, truth and reason. In M. Hollis & S. Lukes (Eds.), *Rationality and Relativism*.

Hacking, I. (2004). *Historical Ontology*. Harvard University Press.

Hacking, I. (2009). *Scientific Reason*. NTU Press.

Hacking, I. (2012). "Language, Truth and Reason" 30years later. *Studies in History and Philosophy of Science Part A*, *43*(4), 599–609. http://doi.org/10.1016/j.shpsa.2012.07.002

Hanna, J. H., Saha, K., & Jaenisch, R. (2010). Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell*, *143*(4), 508–25. http://doi.org/10.1016/j.cell.2010.10.008

Heijmans, B. T., & Mill, J. (2012). Commentary: The seven plagues of epigenetic epidemiology. *International Journal of Epidemiology*, *41*(1), 74–8. http://doi.org/10.1093/ije/dyr225

Hill, D. P., Smith, B., McAndrews-Hill, M. S., & Blake, J. A. (2008). Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, *9 Suppl 5*(Suppl 5), S2. http://doi.org/10.1186/1471-2105-9-S5-S2

Hirsch, A. (2009). Guest Column: A New Kind of Big Science. *The New York Times*.

Hoehndorf, R., Haendel, M., Stevens, R., & Rebholz-Schuhmann, D. (2014). Thematic series on biomedical ontologies in JBMS: challenges and new directions. *Journal of Biomedical Semantics*, *5*(1), 15. http://doi.org/10.1186/2041-1480-5-15

Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V, Chin, L., & Garraway, L. A. (2013). Highly recurrent TERT promoter mutations in human melanoma. *Science (New York, N.Y.)*, *339*(6122), 957–9. http://doi.org/10.1126/science.1229259

Hunter, L. (2002). Ontologies for programs, not people. *Genome Biology*, *3*(6), interactions1002.1–interactions1002.2. http://doi.org/10.1186/gb-2002-3-6-interactions1002

Husserl, E. (n.d.). *Logische Untersuchungen*. Niemeyer, M.

Jablonka, E., Lamb, M. J., & Zeligowski, A. (2014). *Evolution in Four Dimensions, revised edition: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. MIT Press.

Jacob, C. (2006). *The Sovereign Map: Theoretical Approaches in Cartography Throughout History*. University of Chicago Press.

Jacob, F. (1988/1995). *The Statue Within: An Autobiography*. CSHL Press

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, *3*, 318–56.

Kaczmarek, A., Vandenabeele, P., & Krysko, D. V. (2013). Necroptosis: the release of damage-associated molecular patterns and its physiological relevance. *Immunity*, *38*(2), 209–23. http://doi.org/10.1016/j.immuni.2013.02.003

Kadane, J. B., & Seidenfeld, T. (1990). Randomization in a bayesian perspective. *Journal of Statistical Planning and Inference*, *25*(3), 329–345. http://doi.org/10.1016/0378-3758(90)90080-E

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation* (pp. 410–505). University of Minnesota Press.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, *3*(3), 262–267. http://doi.org/10.1177/2043820613513388

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. http://doi.org/10.1177/2053951714528481

Kohler, R. E. (1994). *Lords of the Fly: Drosophila Genetics and the Experimental Life*. University of Chicago Press.

Köhler, S., Bauer, S., Mungall, C. J., Carletti, G., Smith, C. L., Schofield, P., … Robinson, P. N. (2011). Improving ontologies by automatic reasoning and evaluation of logical definitions. *BMC Bioinformatics*, *12*(1), 418. http://doi.org/10.1186/1471-2105-12-418

Kripke, S. A. (1980). *Naming and Necessity*. Harvard University Press.

Kuhn, T. S. (2012). *The Structure of Scientific Revolutions: 50th Anniversary Edition*. University of Chicago Press.

Kuperstein, I., Bonnet, E., Nguyen, H.-A., Cohen, D., Viara, E., Grieco, L., … Zinovyev, A. (2015). Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*, *4*, e160. http://doi.org/10.1038/oncsis.2015.19

Landecker, H., & Panofsky, A. (2013). From Social Structure to Gene Regulation, and Back: A Critical Introduction to Environmental Epigenetics for Sociology. *Annual Review of Sociology*, *39*(1), 333–357. http://doi.org/10.1146/annurev-soc-071312-145707

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–8. http://doi.org/10.1038/nature12213

Leonelli, S. (2008). Bio-ontologies as Tools for Integration in Biology. *Biological Theory*, *3*(1), 7–11. http://doi.org/10.1162/biot.2008.3.1.7

Leonelli, S. (2009). On the locality of data and claims about phenomena. *Philosophy of Science*, *76*(5), 737–749. Retrieved from http://philpapers.org/rec/LEOOTL

Leonelli, S. (2012a). Classificatory Theory in Biology. *Biological Theory*, *7*(4), 338–345. http://doi.org/10.1007/s13752-012-0049-z

Leonelli, S. (2012b). Classificatory Theory in Biology. *Biological Theory*, *7*(4), 338–345. http://doi.org/10.1007/s13752-012-0049-z

Leonelli, S. (2012c). Classificatory Theory in Data-intensive Science: The Case of Open Biomedical Ontologies. *International Studies in the Philosophy of Science*, *26*(1), 47–65. http://doi.org/10.1080/02698595.2012.653119

Leonelli, S. (2013). Integrating data to acquire new knowledge: Three modes of integration in plant science. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*(4), 503–514. http://doi.org/10.1016/j.shpsc.2013.03.020

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data & Society*, *1*(1), 2053951714534395. http://doi.org/10.1177/2053951714534395

Leuty, R. (2015). New "Google Maps for health" drives California to precision medicine lead. *San Francisco Business Times*.

Lewis, S. E. (2005). Gene Ontology: looking backwards and forwards. *Genome Biology*, *6*(1), 103. http://doi.org/10.1186/gb-2004-6-1-103

Lewontin, R. C. (1970). The Units of Selection. *Annual Review of Ecology and Systematics*, *1*(1), 1–18. http://doi.org/10.1146/annurev.es.01.110170.000245

Lewontin, R. C. (2001). *It Ain't Necessarily So: The Dream of the Human Genome and Other Illusions*. New York Review of Books.

Linnaeus, C. (n.d.). *Linnaeus' Philosophia Botanica*. OUP Oxford.

Linnaeus, C. (1751). *Philosophia botanica in qua explicantur fundamenta botanica: cum definitionibus partium, exemplis terminorum, observationibus rariorum : adjectis figuris æneis*. Apud Godofr. Kiesewetter.

Liu, L., & Özsu, T. (2009). *Encyclopedia of Database Systems*. Springer US.

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine*, *40*(4), 346–58.

Machamer, P. K., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Matthewson, J., & Weisberg, M. (2008). The structure of tradeoffs in model building. *Synthese*, *170*(1), 169–190. http://doi.org/10.1007/s11229-008-9366-y

Matthewson, J., & Weisberg, M. (2009). The structure of tradeoffs in model building. *Synthese*, *170*(1), 169–190. http://doi.org/10.1007/s11229-008-9366-y

Mayer-Schonberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. Hodder & Stoughton.

Maynard, P. (2005). *Drawing Distinctions: The Varieties of Graphic Expression*. Cornell University Press.

Mayr, E. (1961). Cause and Effect in Biology: Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science*, *134*(3489), 1501–1506. http://doi.org/10.1126/science.134.3489.1501

Mayr, E. (1982). *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Harvard University Press.

Mayr, E. (1998). *This is Biology: The Science of the Living World*. Harvard University Press.

Meloni, M., & Testa, G. (2014). Scrutinizing the epigenetics revolution, *9*(August), 1–26. http://doi.org/10.1057/biosoc.2014.22

Mill, J., & Heijmans, B. T. (2013). From promises to practical strategies in epigenetic epidemiology. *Nature Reviews. Genetics*, *14*(8), 585–94. http://doi.org/10.1038/nrg3405

Morange, M. (2000). Gene function. *Comptes Rendus de l'Académie Des Sciences - Series III - Sciences de La Vie*, *323*(12), 1147–1153. http://doi.org/10.1016/S0764-4469(00)01264-6

Morange, M. (2006). Post-genomics, between reduction and emergence. *Synthese*, *151*(3), 355–360. http://doi.org/10.1007/s11229-006-9029-9

Morange, M., & Cobb, M. (2000). *A History of Molecular Biology*. Harvard University Press.

Morgan, M. S. (1999). Learning from models. In M. S. Morgan & M. Morrison (Eds.), *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press. Retrieved from http://eprints.lse.ac.uk/6449/

Morgan, M. S., & Morrison, M. (1999). *Models as Mediators: Perspectives on Natural and Social Science*.

Morrison, M. (2007). *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford University Press.

Müller, H.-M., Kenny, E. E., & Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, *2*(11), e309. http://doi.org/10.1371/journal.pbio.0020309

Müller-Wille, S. (2007). Collection and collation: theory and practice of Linnaean botany. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *38*(3), 541–562. http://doi.org/10.1016/j.shpsc.2007.06.010

Müller-Wille, S., & Charmantier, I. (2012). Natural history and information overload: The case of Linnaeus. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 4–15. http://doi.org/10.1016/j.shpsc.2011.10.021

Müller-Wille, S., & Richmond, M. (*forthcoming*). Revisiting the Origin of Genetics. In S. Müller-Wille & C. Brandt (Eds.), *Heredity Explored: Between Public Domain and Experimental Science, 1850-1930*. MIT Press.

Mullis, K. B. (1990). The Unusual Origin of the Polymerase Chain Reaction. *Scientific American*, *262*(4), 56–65. http://doi.org/10.1038/scientificamerican0490-56

Nagel, T. (1989). *The View From Nowhere*.

Nicholson, D. J. (2012). The concept of mechanism in biology. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 152–163. http://doi.org/10.1016/j.shpsc.2011.05.014

Noble, D. (2008). Claude Bernard, the first systems biologist, and the future of physiology. *Experimental Physiology*, *93*(1), 16–26. http://doi.org/10.1113/expphysiol.2007.038695

O'Malley, M. (2007). Exploratory Experimentation and Scientific Practice: Metagenomics and the Proteorhodopsin Case. *History and Philosophy of the Life Sciences*, *29*(3), 335–358

O'Malley, M. A., & Soyer, O. S. (2012). The roles of integration in molecular systems biology. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 58–68. http://doi.org/10.1016/j.shpsc.2011.10.006

Pace, N. R. (1997). A Molecular View of Microbial Diversity and the Biosphere. *Science*, *276*(5313), 734–740. http://doi.org/10.1126/science.276.5313.734

Parke, E. C. (2014, March 7). Experiments, Simulations, and Epistemic Privilege. *Philosophy of Science*.

Pickstone, J. V. (2001). *Ways of Knowing: A New History of Science, Technology, and Medicine*. University of Chicago Press.

Pigliucci, M. (2013). On the different ways of "doing theory" in biology. *Biological Theory*, *7*(4), 287–297. http://doi.org/10.1007/s13752-012-0047-1

Pigliucci, M., Sterelny, K., & Callebaut, W. (2013). The Meaning of "Theory" in Biology. *Biological Theory*, *7*(4), 285–286. http://doi.org/10.1007/s13752-013-0124-0

Poincaré, H. (1902). *Science and Hypothesis*. Courier Corporation.

Popper, K. (1959). *The Logic of Scientific Discovery* (Vol. 4). Hutchinson & Co.

Quine, W. V. O. (1953). *From a Logical Point of View: 9 Logico-philosophical Essays*. Harvard University Press.

Ratti, E. (2015). Big Data Biology: Between Eliminative Inferences and Exploratory Experiments. *Philosophy of Science*, *82*(2), 198–218.

Renear, A. H., & Palmer, C. L. (2009). Strategic reading, ontologies, and the future of scientific publishing. *Science (New York, N.Y.)*, *325*(5942), 828–32. http://doi.org/10.1126/science.1157784

Rhee, S. Y., Wood, V., Dolinski, K., & Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews. Genetics*, *9*(7), 509–515. http://doi.org/10.1038/nrg2363

Rheinberger, H. J. (2011). Infra-experimentality: From traces to data, from data to patterning facts. *History of Science*, *49*(3), 337–348. http://doi.org/10.1177/007327531104900306

Rheinberger, H.-J. (1997). *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press.

Rheinberger, H.-J. (2008a). What Happened to Molecular Biology? *BioSocieties*, *3*(3), 303–310. http://doi.org/10.1017/S1745855208006212

Rheinberger, H.-J. (2010). *On Historicizing Epistemology: An Essay*. Stanford University Press.

Robinson, P. N., & Bauer, S. (2011). *Introduction to Bio-Ontologies*. CRC Press.

Rossi, P. (1997). *La nascita della scienza moderna in Europa*. Laterza.

Schneider, M. V, & Orchard, S. (2011). *Omics technologies, data and bioinformatics principles*. (B. Mayer, Ed.)*Methods in molecular biology* (Vol. 719). Humana Press. http://doi.org/10.1007/978-1-61779-027-0_1

Schuster, P. (2014). Are computer scientists the sutlers of modern biology?: Bioinformatics is indispensible for progress in molecular life sciences but does not get credit for its contributions. *Complexity*, *19*(4), 10–14. http://doi.org/10.1002/cplx.21501

Selzer, P. M., Marhöfer, R., & Rohwer, A. (2008). *Applied Bioinformatics: An Introduction*. Springer.

Seringhaus, M. R., & Gerstein, M. B. (2007). Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinformatics*, *8*(1), 17. http://doi.org/10.1186/1471-2105-8-17

Shen, H. (2013). Stem cells mimic human brain. *Nature*. http://doi.org/10.1038/nature.2013.13617

Sloan, P. R. (2011). *Creating a Physical Biology: The Three-Man Paper and Early Molecular Biology*. University of Chicago Press.

Smith, B. (1998). Basic Concepts of Formal Ontology.

Smith, B. (2003). Ontology. In L. Floridi (Ed.), *Blackwell Guide to the Philosophy of Computing and Information* (pp. 155 – 166). Oxford: Blackwell

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., … Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, *25*(11), 1251–5. http://doi.org/10.1038/nbt1346

Smith, B., & Ceusters, W. (2010). Ontological realism: A methodology for coordinated evolution of scientific ontologies. *Applied Ontology*, *5*(3-4), 139–188. http://doi.org/10.3233/AO-2010-0079

Stamatoyannopoulos, J. A. (2012). What does our genome encode? *Genome Research*, *22*(9), 1602–11. http://doi.org/10.1101/gr.146506.112

Stanford, P. K. (2015). Unconceived alternatives and conservatism in science: the impact of professionalization, peer-review, and Big Science. *Synthese*. http://doi.org/10.1007/s11229-015-0856-4

Steinle, F. (1997). Entering new fields: Exploratory uses of experimentation. *Philosophy of Science*, *64*(4), 74

Stevens, H. (2013). *Life Out of Sequence: A Data-Driven History of Bioinformatics* (Vol. 4). University of Chicago Press.

Strasser, B. J. (2006). Collecting and Experimenting; The Moral Economies of Biological Research, 1960s- 1980s. In S. de Chadarevian & H.-J. Rheinberger (Eds.), *History and Epistemology of Molecular Biology and Beyond*. Max-Planck Institute for the History of Science.

Strasser, B. J. (2008). GenBank--Natural history in the 21st Century? *Science*, *322*(5901), 537–8. http://doi.org/10.1126/science.1163399

Strasser, B. J. (2010). 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*(4), 623–60. http://doi.org/10.1007/s10739-009-9221-0

Strasser, B. J. (2012a). Collecting Nature : *Osiris*, *27*, 303–340.

Strasser, B. J. (2012b). Collecting Nature : Practices, Styles, and Narratives. *Osiris*, *27*(1), 303–340. http://doi.org/10.1086/667832

Strasser, B. J. (2012c). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 85–7. http://doi.org/10.1016/j.shpsc.2011.10.009

Strasser, B. J. (2014). *Biomedicine: Meanings, assumptions, and possible futures*. Retrieved from http://www.swir.ch/images/stories/pdf/en/SWIR_1_2014_Biomedicine.pdf

Strasser, B. J., & Chadarevian, S. de. (2011). The comparative and the exemplary: revisiting the early history of molecular biology.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719–24. http://doi.org/10.1038/nature07943

Suárez, M. (2003). Scientific representation: against similarity and isomorphism. *International Studies in the Philosophy of Science*, *17*(3), 225–244.

Swoyer, C. (1982). The nature of natural laws. *Australasian Journal of Philosophy*, *60*(3), 203–223. http://doi.org/10.1080/00048408212340641

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*(3), 449–508. http://doi.org/10.1007/BF00499820

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, *126*(4), 663–76. http://doi.org/10.1016/j.cell.2006.07.024

Vandenabeele, P., Galluzzi, L., Vanden Berghe, T., & Kroemer, G. (2010). Molecular mechanisms of necroptosis: an ordered cellular explosion. *Nature Reviews. Molecular Cell Biology*, *11*(10), 700–14. http://doi.org/10.1038/nrm2970

Vettel, E. J. (2008). *Biotech: The Countercultural Origins of an Industry*. University of Pennsylvania Press.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science (New York, N.Y.)*, *339*(6127), 1546–58. http://doi.org/10.1126/science.1235122

Waddington, C. H. (1942). The epigenotype. *Endeavour*, *1*(18).

Waddington, C. H. (1968). Towards a Theoretical Biology. *Nature*, *218*(5141), 525–527. http://doi.org/10.1038/218525a0

Waddington, C. H. (2012). The epigenotype. 1942. *International Journal of Epidemiology*, *41*(1), 10–3. http://doi.org/10.1093/ije/dyr184

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., & Lewis, S. E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biology*, *7*(11), e1000247. http://doi.org/10.1371/journal.pbio.1000247

Waters, K. (2007). The nature and context of exploratory experimentation: an introduction to three case studies of exploratory research. *History and Philosophy of the Life Sciences*, *29*(3), 275–84.

Weber, M. (2005). *Philosophy of Experimental Biology*. Cambridge University Press.

Weinberg, R. (2010). Point: Hypotheses first. *Nature*, *464*(7289), 678. http://doi.org/10.1038/464678a

Weisberg, M. (2007b). Three Kinds of Idealization. *The Journal of Philosophy*, *104*(12), 639–659.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. OUP USA.

Williams, B. (2002). *Truth and Truthfulness: An Essay in Genealogy*. Princeton University Press.

Wilson, E. B. (1901). Aims and Methods of Study in Natural History. *Science*, *13*(314), 14–23. http://doi.org/10.1126/science.13.314.14

Wilson, E. O. (1994). *Naturalist*. Island Press.

Wimsatt, W. C. (2007). *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Harvard University Press.

Winther, R. G. (2011). *Part-whole science*. *Synthese* (Vol. 178). http://doi.org/10.1007/s11229-009-9647-0

Winther, R. G. (2012a). Interweaving categories: Styles, paradigms, and models. *Studies in History and Philosophy of Science Part A*, *43*(4), 628–639. http://doi.org/10.1016/j.shpsa.2012.07.005

Winther, R. G. (2012b). Mathematical modeling in biology: philosophy and pragmatics. *Frontiers in Plant Science*, *3*, 102. http://doi.org/10.3389/fpls.2012.00102

Winther, R. G. (2014). Mapping Kinds in GIS and Cartography. In C. Kendig (Ed.), *Natural Kinds and Classification in Scientific Practice*.

Wittkop, T., TerAvest, E., Evani, U. S., Fleisch, K. M., Berman, A. E., Powell, C., … Mooney, S. D. (2013). STOP using just GO: a multi-ontology hypothesis generation tool for high throughput experimentation. *BMC Bioinformatics*, *14*(1), 53. http://doi.org/10.1186/1471-2105-14-53

Woolf, S. H. (2008). The meaning of translational research and why it matters. *JAMA*, *299*(2), 211–3. http://doi.org/10.1001/jama.2007.26

Yaffe, M. B. (2013). The scientific drunk and the lamppost: massive sequencing efforts in cancer discovery and treatment. *Science Signaling*, *6*(269), pe13. http://doi.org/10.1126/scisignal.2003684

Yang, J., Chen, L., Kong, X., Huang, T., & Cai, Y.-D. (2014). Analysis of tumor suppressor genes based on gene ontology and the KEGG pathway. *PloS One*, *9*(9), e107202. http://doi.org/10.1371/journal.pone.0107202