PhD degree in Foundations of the Life Sciences and their Ethical Consequences

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

Settore disciplinare: M-FIL02

THE CONTEXT OF DISCOVERY OF DATA-DRIVEN BIOLOGY

Emanuele Ratti

Matricola n. R09884

Supervisor: Professor Giovanni Boniolo, Dr Marco Nathan

Added supervisor: Professor Michael Weisberg

Anno accademico 2014-2015

ABSTRACT

My PhD dissertation aims (1) at reconstructing the structure of the context of discovery of 'data-driven' (big data, data intensive) biology and (2) at comparing it to traditional molecular approaches. Within the current debate in philosophy of science, 'traditional approaches' in molecular biology should be understood as the discovery and heuristics strategies identified by mechanistic philosophers such as Carl Craver and Lindley Darden. Therefore, key questions of my thesis are: *what is the structure of discovery of data-driven biology? Is data-driven biology methodology different from traditional molecular approaches?*

The reason for doing such an analysis comes from a recent controversy among biologists. In particular, sides disagree on whether high throughput sequencing technologies are stimulating the development of a new scientific method somehow irreducible to traditional approaches. I will try to disentangle the debate by reconstructing and comparing data-driven and traditional methodologies. The dissertation is composed of five chapters.

The first chapter deals with methodological issues. How do I compare data-driven and traditional molecular biology structures of discovery? Mechanistic philosophers have extensively characterized the discovery structure of traditional molecular biology. However, there is not such an analysis for data-driven biology. In order to do this, I will critically revise the discovery/justification distinction. The debate on discovery/justification has provided valuable tools on how discovery strategies might be conceived, and it is clearly one of the main forefathers of recent philosophical discussions on scientific methodologies in biology and physics.

In Chapter 2 I shall to try to infer a full-fledged account of discovery for data-driven biology by means of the philosophical tools developed in Chapter 1. This analysis will be done in parallel to the investigation of key examples of data-driven biology, namely genome-wide association studies and cancer genomics. In Chapter 3 I analyze the epistemic strategies enabled by biological databases in data-driven biology. In Chapter 4, I will show how the discovery structure of 'traditional molecular biology' can be more efficiently rephrased through the same theoretical framework that I use to characterize data-driven biology.

Since data-driven and traditional molecular biology seem to adopt the same discovery structure, one might consider the controversy motivating my research ill posed. However, in Chapter 5 I shall argue that there is still a valuable reason of disagreement between the sides. Actually, data-driven and traditional molecular biology endorse different cognitive values, which provide the criteria for evaluating models and findings as adequate or not. Here one might say that, although the structures of discovery (i.e. how

reasoning and experimental strategies are structured and depend on each other) of the two sides are the same, the *contexts of discovery* (i.e. the set of both reasoning/experimental strategies *and* epistemic values/background assumptions that motivate discovery) are different. Therefore, in this last chapter I shall pinpoint the cognitive values behind traditional and data-driven biology, and how these commitments stimulate the heated disagreement motivating my research.

ACKNOWLEDGMENTS

I would like to thank the faculty and the colleagues of FOLSATEC. This program has been for me a source of intellectual enlightenment and personal growth that it could have been hardly found elsewhere. In particular I would like to thank Federico Boem, who is the person that not only introduced me to the philosophy of biology, but also that taught me the meaning of 'stimulating exchange of ideas' on any topic. A particular mention goes to Pierre-Luc Germain (more a mentor than a colleague), who challenged my ideas in a very constructive way and put a lot of efforts in helping me to grow intellectually. A special thank goes to my supervisor Giovanni Boniolo, both for his timely comments on my work and for his efforts in constructing the incredible and stimulating intellectual environment that FOLSATEC represents. I am thankful to my co-supervisor Marco Nathan, for his efforts in putting my philosophical arguments in the right direction. I am in debt also with Michael Weisberg (my external supervisor) for indispensable discussions on the big data phenomenon that drove my dissertation in the right direction. I am grateful to Francesca Ciccarelli for hosting me in her lab, and for having taught me many important things on genomics and on science in general. Last (but not least), I am extremely thankful to David Teira for his indispensable advises on philosophy, life and everything. I would like to thank also Veronica Viscardi and Francesca Fiore (coordinators of SEMM programme) for their laudable help in these four years.

Finally – and most important – I am indebted to people outside the academic environment. My parents encouraged me during the last 10 years to cultivate my own interests: I will not forget this. My girlfriend Jennifer supported me in sad and happy moments of my intellectual life: I hope I will be able in the future to pay off my debts to you.

INTRODUCTION

AN EPISTEMOLOGICAL CONTROVERSY BETWEEN TRADITION AND INNOVATION

These are great times for being a philosopher of science, especially a philosopher of biology and/or medicine. As in several scientific fields, in molecular biology things are changing at a hectic pace. This is not to say that there is progress, but just that there is something that is changing towards unknown directions. Clearly, there has been acceleration in biology after the first releases of the Human Genome Project (Lander et al 2001) and its competitors (Venter et al 2001). The Human Genome Project (HGP) has been an incredible catalyst of transformation for biology in many respects (Hood and Rowen 2013). In particular, HGP drove the development of sequencing technologies (e.g. Next-Generation Sequencing) that eventually are stimulating a maniacal focus on data production.

The amount of data produced by biology in the new century has reached a psychological threshold and this has created tensions among biologists. In this dissertation, I will try to disentangle a methodological dispute related to how biologists perceive this focus on data production. In particular, the controversy is over the nature of contemporary discovery strategies based on data intensive approaches and sequencing technologies stemming from the post-HGP era (the so-called 'data-driven approach[1]'). Factions in this debate agree on the idea that there are a 'traditional' and a 'novel' way of doing molecular biology, but disagree on which approach is the best for molecular biology. Some argue (Weinberg 2010, Alberts 2012) that 'data-driven' molecular biology makes use of a method that is not properly scientific, and that this methodology undermines the survival of a 'traditional' way of doing molecular biology. Others (Golub 2010; Garraway and Lander 2013) support the idea that a 'data-driven' approach is genuinely scientific, and that it can provide insights that traditional approaches cannot. Here, I will provide an epistemological analysis of both approaches in order to understand where exactly the disagreement lies. Therefore, the aim of this project is to characterize the context of discovery of contemporary molecular biology.

[1] I do not think that this label is particularly appropriate. I share these doubts with other intellectuals and practitioners (Smalheiser 2002; Callebaut 2012; Strasser 2012; Leonelli 2012a; 2012b). However, since the received view on this kind of biology is to call it 'data-driven' I will use this label throughout this dissertation anyway.

Before going into the detail of this controversy and the questions I would like to answer, I shall make a short introduction to disentangle certain terms like 'big data', 'big science' and 'data-driven', in order to mark the boundaries of the biology I will talk about.

Big Data, big science and data-driven approaches

Recently scientists, policy-makers, economists and sociologists have started debating the issue of 'big data'. However, there are different interpretations of what 'big data' actually means. It seems that 'big data' is a new label describing a paradigm shift in how to organize multiple aspects of contemporary ICTs-based societies, ranging from scientific research to societal issues. A common idea is that nowadays information and communication technologies (ICTs) produce unprecedented massive amounts of data on a daily basis. Somehow this restless production of data is changing the way we organize society, encompassing all of its aspects. There is also interplay between technological and cultural (even mythological) aspects (boyd and Crawford 2012). Technological aspects should be identified with the possibilities of the ever-increasing computational power of contemporary ICTs. Mythological and cultural feelings rest on the idea "that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy" (2012, p 3). Therefore, people – from the computer scientist to the sociologist - are arguing in favour of an on-going revolution encompassing all aspects of human society due to the 'big data' phenomenon.

In the context of scientific research sometimes people superimpose 'big data' to 'data-driven science'. I do not think it is accurate. As I am going to show, not all big data scientific projects are strictly driven by data (although the aspect related to the data 'deluge' is prominently highlighted), but rather by hypotheses or by other types of aims. Let us clarify the big data phenomenon in the context of scientific research.

First of all, in science 'big data' can be identified with the phenomenon of 'big science', although Big Science does not necessarily imply the production of data usually associated to Big Data. For instance, The Manhattan Project involved more than 130,000 employees as well as 2 billions dollars of investment, but it was not a 'big data' project, at least not directly.

According to Eddy (2013) there are three kinds of big science: the big experiment, the big map, and the leading wedge.

The big experiment is, by definition, a specific experiment (therefore guided by a single and precise hypothesis) requiring a large-scale community effort. A good case in point is the detection of the Higgs Boson at the Large Hadron Collider in Geneva in 2012.

The idea is that, in order to test the theoretical model elucidating the features of the Boson particle, one requires a particularly complex instrument to do a specific experiment. In this case, the big experiment involves the design of thousands of small experiments, the analysis of 25 petabytes of data (as to 2012) and a facility (the Large Hadron Collider) where nearly 10,000 people work. Actually, the Large Hadron Collider is the most expensive and largest scientific infrastructure ever built. The reader might understand why the big scientific experiment of the Higgs Boson fits very well the requirements of the big data phenomenon. Although these experiments generate a massive amount of data, they are clearly driven by a hypothesis (the theoretical model of the Higgs Boson). Experiments are designed in order to obtain a specific effect, i.e. the Higgs Boson itself.

Next, a big map "is a data resource – comprehensive, complete, closed ended – to be used by multiple groups, over a long time, for multiple purposes" (Eddy 2013, p R261). Biology is a discipline more prone to maps rather than big experiments. A paradigmatic example of map in big data biology is HGP (Lander et al 2001; Venter et al 2001). This project had (and it is having) a huge impact on biology and medicine. In a nutshell, HGP is used as a map to orient research (e.g. to locate genes, to compare nucleotide variants, etc). This map, although in principle it is not indispensable, is crucial in practice.

Finally, a leading wedge is a project designed as a massed technology development effort. Clearly Eddy refers to leading wedges as different in type from big experiments and maps. However, while maps and big experiments are mutually exclusive, leading wedges *and* maps (or leading wedges *and* big experiments) are not. Actually, both the big experiment and the big map imply somehow a leading wedge. HGP involved the massive development of sequencing technologies while the Higgs boson experiment required an incredibly advanced technological infrastructure (the Large Hadron Collider) as well as a great development of computational power to deal with 25 petabytes of data. Big experiments and maps fuel technological ruptures. Some argue that there are also big science project that are leading wedges *tout court*. For instance, recent debates on the Human Brain Project (HBP, see http://www.neurofuture.eu/) raise the issue of whether the HBP might be conceived as a large-scale scientific project with only a technological rationale. Eddy stresses this point also for the Brain Activity Map (BAM) that "is not a map, nor an experiment; it is all leading wedges" (Eddy 2013, p R261).

To my knowledge, there are no big experiments in biology. Better: There are no experiments in biology that are even comparable in size to experiments such as the Higgs Boson briefly explained above. Eddy emphasizes this by saying that "[w]e feel a strong temptation to spin all big science projects in biology as big experiments, whereas the

complexity of biology (…) is such that big experiments are nearly nonexisting in our field" (Eddy 2013, p R261). As Eddy further stresses, HGP was sometimes spun as revolutionary for *understanding* the human genome. But this was wrong, because it just provided the sequence of the human genome without explaining much how the human genome works – and yet HGP as a resource is essential.

But if there are no big experiments in biology, is big data biology just a bunch of maps? Not really. I think that Eddy misses an important type of big science project. This type of project is called the *data-driven project*[2]. Data-driven projects are neither big experiments, nor maps, but they might imply leading wedges. A good candidate for being both data-driven and leading wedges are many studies fuelled by The Cancer Genome Atlas, which clearly has boosted sequencing technologies towards a great improvement. Data-driven projects are not big experiments because they are not driven by a specific hypothesis as the big experiments are. In the context of big data biology, a data-driven project is a project that generates a massive amount of data without a specific hypothesis in mind. So far, it is similar to the map: Generating data without somehow evaluating claims. Yet, in another sense data-driven projects are not like maps. Data-driven projects have an aim of their own, unlike maps that are just designed to aid the biological community in a way or another. Maps are tools to orient research, and aid biologists by offering a basis of comparison in several respects. If maps "are about enabling small science" (Eddy 2013, p R261) and they are not the science itself, data-drive projects are instances of big science, but they have the same aims of small science. Data-driven projects generate big genomic data sets and by means of comparison with either other data sets or maps (in the sense explained above) are able to detect patterns that traditional molecular biological approaches would not be able to detect – or so supporters of data-driven projects say. *In this dissertation I shall be focused especially on data-driven projects*. However I will extensively also talk about maps, since maps are involved in the heuristic strategies used in the discovery procedures of data-driven molecular biology. Let us now see the methodological controversy over data-driven projects that motivates the present work

Data-driven molecular biology: a methodological controversy

In the last few years, molecular biology has become more and more entrenched with sequencing technologies. This is a consequence of the expansion and hype of so-called –*omics* technologies. There is a clear tendency to use these technologies to generate

[2]Some calls data-driven projects 'hypothesis-free' although this is not an entirely appropriate label, as I will show

bigger and bigger data sets, and then to discover significant patterns instead of starting from a single general guess and to design a plethora of experiments to refine the initial claim. However, there has been also a tendency to contrast this new fashion. For instance Weinberg (2010), who identifies the method of molecular biology as being 'hypothesis-driven', claims that 'data intensive[3]' studies in biology have not, so far, shown the same success as the 'traditional' methodological paradigm of starting from hypotheses and then applying a reductionist approach of studying parts of a system in isolation. Weinberg is claiming that a biological study, to be scientific, should start necessarily from hypotheses. Data intensive studies, Weinberg argues, are the new fashion just because there is the misleading perception that traditional studies cannot, in principle, map the complexity coming out of studies using sequencing technologies. Todd Golub (2010) on the contrary argues that the 'data-driven' approach can lead to discoveries that traditional studies (identified as 'hypothesis-driven') could not reach. Golub is particularly diplomatic, and he runs with the hare and hunts with the hounds, especially when he says that "[a]lthough hypothesis-driven, experimental research should remain central to the field, unbiased surveys of cancer genomes afford an unprecedented opportunity to generate new ideas" (2010, p 679), but clearly he stresses the point that "[t]his large-scale, data-harvesting approach to biological research has significant advantages over conventional, experimental methods" (2010 p 679). Weinberg and Golub's contributions have been published in the same issue of *Nature* with the paradigmatic titles of, respectively, *Point: Hypothesis First* and *Counterpoint: Data First*. But this is just the tip of the iceberg: The disagreement is pervasive. Indeed, there are different levels in this controversy.

First, there is a controversy over the notion of proper scientific method and the best mode of research. As stressed in O'Malley *et al.* (2009), it is conventional for scientists and philosophers to claim that one of the main activities that can successfully demarcates science from non-science is hypothesis testing. This clearly has led to an oversimplification of the practices of science, thereby ignoring all other inductive and exploratory activities that have populated in a way or another the history of science. Somehow this was also reflected in the guidelines of funding agencies (O'Malley et al 2009) where the 'hypothesis-driven' mode has been prominent for several decades. However, in contemporary molecular biology the accumulation of data and the development sequencing technologies points to other – and different – modes of research. For instance, O'Malley and colleagues report the example of the NIH. While to some degree this funding agency recognizes also a place for exploratory experiments, the focus on hypothesis testing is pervasive. But this way of conceiving science can have some puzzling consequences. For instance Gannon (2009) publishes a parody of a peer review funding process where Charles Darwin's project is rejected because of the lack of a

[3] In this work 'data-driven' and 'data-intensive' are synonymous

clear hypothesis guiding his research. On the other hand, there is a tendency to be overly enthusiastic about data intensive enterprises (Yaffe 2013). Therefore, an aspect of the controversy deals with the right scientific method for biology, with people like Weinberg claiming that the 'hypothesis-driven' and reductionist approach is the right one for biology, and scientists like Golub claiming that only a data-driven approach can in principle make sense of biological complexity.

The second level is on the opposition between 'data-driven' and 'hypothesis-driven'. But what is exactly a 'data-driven approach'? As far as I have understood from the received view on this matter, an approach is data-driven – as opposed to hypothesis-driven – if it designs an experiment merely to generate relevant data and not to obtain a specific effect in order to test (*directly*) a specific causal claim. In a sense, data-driven projects are devoted to the generation of hypotheses. Biologists like Eddy (2013) or Weinberg (2010) think that molecular biology is mainly a matter of testing claims, and a scientific research without hypotheses is blind and misguided from the very beginning. Others think that a hypothesis-free or data-driven approach is in principle unbiased and therefore it has several advantages over approaches driven by guesses. In particular in the context of genome-wide association studies (GWASs), these types of screenings are seen as eminently unbiased so that they can lead to genuine new discoveries. For instance Brookfield (2010) argues that the "GWAS approach is hypothesis-free, in that it looks at very many SNPs simultaneously rather than focusing on loci whose biology suggests that a causal relationship to the disease is likely" (p 4). Another example is the commentary by Guessous and colleagues (2009), where they claim that "[t]he conventional GWAS approach is a hypothesis-free, systematic search of tagging single nucleotide polymorphisms (SNPs) across the genome to identify novel associations with common diseases" (p 1). Finally, Yeo writes an editorial (2011) and he says that "because of its 'hypothesis free' nature, the power of GWAS lies in uncovering potentially new biology that would not have been possible using a candidate gene approach" (p 1), where a 'candidate gene approach' is a typical hypothesis-driven mode of research. Therefore, on the one hand 'traditional' biologists think that proper science should be done with hypotheses, while 'data-driven' practitioners claim that a data-driven approach could overcome the limits of the old method.

Third, there is the 'reductionist versus non-reductionist' quarrel. 'Reductionist' should not be understood in the technical philosophical meaning, but rather in a more intuitive fashion. Especially in molecular oncology (the main case study in this work), the quest for cancer genes can be approached in two ways. Either we do a sort of 'trial-and-error' approach, and we test genes to see their effect on cancer phenotypes, or we establish some criteria that a gene should met to be of interest, and by means of sequencing technology we try to see whether all genes - simultaneously – meets these

criteria. The latter is the idea that data-driven molecular biology can provide a system-level view of biological systems. Vogelstein and colleagues (2013, 1554) describe this idea by means of a metaphor saying that if a jungle appear chaotic at ground level, one can distinguish several interesting patterns by taking a view from 30,000 foots. The '30,000-foot view' is the system level view of data-driven biology. Please note that the 'system-level' view versus reductionist view is not an ontological controversy, at least not here. It is purely methodological, namely it is the question "What is the best approach to attack the complexity of biological systems?". Therefore, data-driven biologists claim that their approach is the most effective, while 'traditional' biologists let 50 years of successes of reductionist approaches speak for themselves.

There is also another aspect of the controversy. 'Data-driven' molecular biology is mostly organized around big consortia. For instance, to do a GWAS more is needed than just a single lab: Funding data-driven projects is particularly demanding. Therefore the controversy is also about which projects we should fund. For instance Alberts (2012) thinks that small lab-based science is somehow threatened by the emergence of big consortia.

To sum up, there is a controversy between a well-established tradition of discovery strategies of molecular biology called 'hypothesis-driven' that I shall call 'traditional molecular biology', and an emerging community of scientists that identify themselves as 'data-driven', 'hypothesis-free', 'systemic' (that is different from 'systems biology') molecular biologists that I will call, for simplicity, 'data-driven' molecular biology. In this dissertation I would like to answer to the following questions:

Which are the epistemological differences between 'traditional' and data-driven molecular biology?

Is this controversy misplaced and misguided?

If not, where does exactly the disagreement lie?

Recently, a proposal has been advanced. Instead of being in opposition, data-driven and hypothesis-driven are complementary and hybridized (Strasser 2011; Smallheiser 2002; O'Malley and Soyer 2012; Keating and Cambrosio 2012). The problem with these proposals is that they do not explain, from an epistemological point of view, how these two approaches are complementary. Therefore in this dissertation I will analyse the discovery strategies of both data-driven and traditional molecular biology in order to highlight similarities and differences.

Structure of the thesis

The dissertation is structured as follows. Since I will analyse from a philosophical point of view discovery strategies, in Chapter 1 I will scrutinize the debate in philosophy of science over methodologies of discovery and I will try to elaborate a minimal framework of how discovery in science is pursued. To do this, I will go through the debate fuelled by Reichenbach and its controversial distinction between the context of discovery and the context of justification. The debate stems from this distinction has continued for several years, leading to the present mechanistic philosophy. I will then come out with my framework of scientific discovery, combining virtues of different positions emerged in the debate.

In Chapter 2 I will analyse data-driven molecular biology strategies of discovery in light of the framework elaborated in Chapter 1. I will analyse especially case studies from cancer genomics and GWASs.

Chapter 3 should be conceived as complementing Chapter 2. In Chapter 2, I shall point to the important epistemic role that biological databases play in contemporary biology. In Chapter 3 I distinguish two important epistemic uses of biological databases. In a first sense, databases are evidence-enhancers, in the sense that they aid biologists in identifying biological phenomena and in elaborating claims about phenomena. In a second sense, databases are 'mined' with theoretical aims in mind.

In Chapter 4 I shall focus on 'traditional' discovery strategies in molecular biology, especially in molecular oncology. The orienteering tool here is the idea that discovery strategies in molecular biology are attempts to build mechanistic descriptions of biological phenomena. Therefore, I will draw from the literature of the so-called mechanistic philosophy (Bechtel and Richardson 2010; Darden 2006, Craver and Darden 2013) to identify the main epistemic features of molecular biological practices, and I will then compare these with the framework elaborated in Chapter 1.

In Chapter 5, I will compare the two approaches and I will claim that the main reason motivating the heated disagreement lies in the endorsement of different cognitive value of a special kind.

CHAPTER 1

A TRIPARTITE FRAMEWORK FOR SCIENTIFIC DISCOVERY

CHAPTER ABSTRACT

In this chapter, I discuss what scientific discovery is, how it is structured, and which criteria might influence its development. These conceptual tools will be applied to both data-driven and traditional molecular biology to grasp the discovery strategies underpinning these enterprises. I draw these conceptual tools by analyzing in detail the debate about the distinction between the context of discovery and the context of justification, that is arguably the *locus* in the philosophy of science literature where discovery was debated most. I analyze the most influential positions on this topic and argue that scientific discovery is composed of three phases: *hypothesis generation*, *hypothesis prior-assessment* and *hypothesis justification*. Moreover, I list some possible types of inferences used within each phase. At the end, I complement this analysis by introducing the topic of epistemic values showing why we should consider these types of epistemic desiderata when analyzing discovery strategies.

1. INTRODUCTION

In philosophy, the topic of discovery has been extensively analyzed and explored. Some have seen the problem of discovery as a traditional issue in epistemology dealing with the 'justificatory' part of knowledge (viz. 'justified true belief') while others have focused more specifically on scientific discovery. In the latter case, the term 'discovery' has many interpretations (Nickles 1980; Schickore 2014). One might think about 'discovery' as a recognized scientific achievement, or as *the process* leading to a scientific achievement. It can be either an 'ah-a' experience or a comprehensive and successful scientific inquiry. Conceived in a stronger sense, 'discovering' can be identified with 'knowing', while in a weaker sense 'discovering' is just the act of having an idea, implying no evaluation of that idea whatsoever.

In this dissertation, by 'discovery' I mean *the process leading to a scientific achievement*. Therefore, when I said in the Introduction that I want to compare discovery strategies and structures of data-driven and 'traditional' molecular biology, I meant that I want to compare the processes leading to their scientific achievements. In order to do this, I need a framework of scientific discovery that can be used as a basis for the comparison of the two approaches. This does not mean that I have to elaborate a framework able to represent the formal structure of *any* scientific discovery. Rather, what I need is an account that might include a *minimal core* of phases and epistemic moves that any scientist should pass through when 'discovering'. In other words, this would be a framework providing a set of necessary steps and epistemic moves of scientific discovery. With this account in place, I can identify similarities and differences of the two approaches within each phase of scientific discovery.

I have three goals for this chapter. First, I want to show that scientific discovery, understood as the process leading to a scientific achievement, is a procedure constituted by *at least* three phases (hypothesis generation, hypothesis development and evaluation). I call this process of generation-development-evaluation the *tripartite framework of discovery*. In order to make the case for my tripartite framework, I critically scrutinize the debate on the distinction between the context of discovery and the context of justification that is arguably the main debate on scientific discovery in the philosophy of science. Next, I show that within each phase there is an *umbrella* of possible inferences that can be used. Finally, I argue that each inference – depending on the phase where it is applied and on the context to which is embedded in - is guided by a set of preferences that are called 'epistemic desiderata'.

2. THE DISTINCTION BETWEEN THE CONTEXT OF DISCOVERY AND THE CONTEXT OF JUSTIFICATION

In philosophy of science, the topic of discovery, understood as the process leading to a scientific achievement, has been discussed as part of the debate on the distinction between the context of discovery and the context of justification. This distinction stems from Reichenbach's *Experience and Prediction* (1938; 1961) and *The Rise of Scientific Philosophy* (1951). The distinction has been spelled out in many ways, but there is a core in Reichenbach's initial proposal that has been widely discussed and has generated different and conflicting positions.

In outlining the descriptive task of epistemology, Reichenbach (1961) argues that this discipline is interested in internal relations between thoughts. But if epistemology is interested in internal relations between thoughts, then it seems that it should also be

focused in providing descriptions of thinking processes in the way they are actually performed. According to Reichenbach, this is deeply wrong. Epistemology is only interested in the logical interconnections between thoughts, and not in how thoughts are performed. Therefore, epistemology is aimed at constructing "thinking processes in a way in which they ought to occur if they are to be ranged in a consistent system" (Reichenbach 1961, p 4). Whereas psychology describes actual thinking processes as they are actually performed, epistemology rather focuses on a 'justifiable set of operations' between the starting point and the end of a thought processes. In other words, epistemology has to reconstruct the way in which thinking processes are communicated to other individuals, e.g. the way a discovery is presented in a peer-reviewed journal[4]. Hence the difference between psychology and epistemology is the same as the difference between a person's way of finding (for example) a theorem, and the way of demonstrating the theorem itself. In order to better explain the difference, Reichenbach introduces the distinction between the context of discovery (*the way thoughts are actually performed*) and the context of justification *(the way thoughts are rationally reconstructed*). Moreover, in (1951) Reichenbach elaborates a stronger position. Namely, he argues that the scientist who discovers a theory is guided by guesses, and "he cannot name a method by means of which he found the theory" (1951, 230). Reichenbach explicitly says that there is no way to rationally discover a theory or a hypothesis. How a theory is generated is more an empirical question, and it is investigated by disciplines such as history or psychology. Therefore according to Reichenbach one first discovers a theory or a hypothesis in a non-rational way, and then she justifies it.

The core of the distinction then is as follows. In the context of discovery, one generates a hypothesis by tentative guesses, and how exactly this has been done is investigated by empirical disciplines such as history or psychology. In the context of justification, one then tries to justify the generated hypothesis. The way a hypothesis is justified can be subjected to logical inquiry, and this means that only the context of justification is the realm of investigation of philosophy of science because the context of discovery is alogical at best, if not illogical. From this sketch, the distinction of the contexts has been understood in many ways. For instance Hoyningen-Huene (1987; reprinted and further developed in 2006) distinguishes five major ways of understanding the distinction, while Nickles (1980) identifies seven forms of it. Since I am not interested in subtle distinctions, I will divide positions in this debate according to two problems. The first problem is whether a distinction between a phase where a hypothesis is generated and a phase where a hypothesis is evaluated is sufficient to make sense of the complexity of scientific practice. For instance, is it sufficient to say that in actual practice scientists

[4] This is called by Carnap 'rational reconstruction'

generate a complete hypothesis that is then put at test? Second, what is debated is the nature of the first phase, whether it is rational and, if this is the case, whether there are clearly identifiable rules for hypothesis generation. Stemming from these problems, there are two big families of positions about scientific discovery (understood as the process leading to scientific achievement). The first is the family of those who think that there is nothing epistemologically relevant for science in how hypotheses or theories are discovered – for the sake of simplicity I will call this group of ideas *the Popperian family*, because Popper (2002) is arguably the most influential philosopher supporting this position. This family of positions is useful to discuss the issue of whether the process of generating hypotheses can be in principle rational and hence whether can be the subject of a philosophical inquiry. The second set of positions include those who *do* think that discovery is epistemologically relevant for science, and that it can be *at least* rationally reconstructed. Echoing Ronald Giere (Nickles 1980; 1985) I will call this 'family' *friends of discovery*. These positions are useful to investigate whether there are just hypotheses generation and justification, or whether there is something else.

2.1 The Popperian Family

Views belonging to the Popperian family explicitly separate belief-forming procedures from belief-justification procedures by claiming that *how do we form belief (how do scientists generate a hypothesis) has nothing to do with the epistemology of science*. Therefore, the attitude towards discovery as a generative phase of hypotheses is that *discovery procedures are epistemologically irrelevant*. This is because discovery strategies are mainly inductive (or so the Popperian family claims), and induction is unreliable. This position is sometimes developed by saying that, since discovery is irrelevant to epistemology, and epistemology should be logically based, then discovery has nothing to do also with logic. Again, this stemmed from the idea that discovery was seen as an inductive enterprise, and for induction there was not a clear framework, let alone a strict logic. Prominent philosophers supporting this type position are Popper (2002), Braithwaite (1953), Hempel (1966), but also Feigl (1970). Here I just focus on two main positions (Popper and Hempel), that cover what other members of this family have said.

In *Logic of Scientific Discovery* (2002), Popper draws a sharp distinction between the psychology of knowledge (dealing with empirical facts) and the logic of knowledge (dealing only with logical relations). Attempts to construct an inductive logic, Popper says, stems from a misleading belief that confuses psychological and epistemological problems. This idea of identifying logical laws with psychological laws is called by Popper *psychologism* (Popper 2002, p 7).

The idea that the early stage of scientific inquiry - namely the act of conceiving or inventing a theory - can be rationally reconstructed is pervaded by psychologism. Unlike the methods of justification, this early phase cannot be subjected to any kind of logical analysis. Hence, Popper distinguishes sharply between the process of conceiving a new idea (hypothesis or theory), and the methods for examining it logically. The epistemology of science (the logic of knowledge) has to do only with the latter phase. But, even if we admit that the task of epistemology of science is to do the *rationally reconstruction* of the whole process – the steps that have led the scientists to the discovery of new truths – then we should ask what could be in principle reconstructed. Since for Popper hypothesis generation is the "stimulation and release of an inspiration" (2002, p 8), and since inspirations contain creative elements, then there is no method of conceiving hypotheses or theories, because these processes contain an irrational element of creativity that, by definition, cannot be subjected to logical analysis. In Popper's analysis there is no way, not even a very contingent one, through which the phase of hypothesis generation can be somehow subjected to logical analysis. The phase cannot be rationalized in any way.

Hempel's argument (1966, Chapter 2) is close to Popper's, but Hempel provides a stronger motivation for endorsing it. The question he wants to answer is the following: How are hypotheses derived in the first place? A standard answer, Hempel says, is to state that hypotheses are inferred from antecedent collected data by means of 'inductive inferences'. This idea of 'inductive inferences' is, in a particular wave of philosophy of science, a synonym of 'hypothesis generation'. Hempel identifies four stages of an ideal scientific inquiry guided by 'inductive inferences':

(a) Observation in which all facts are recorded
(b) Facts are analysed and classified
(c) Generalizations (theories of hypotheses) are inductively derived
(d) Generalizations are tested

According to Hempel, this scheme has serious problems. The first thing to be noted about (a) is that a scientific investigation "could never get off the ground" (Hempel 1966, p 11) without hypotheses. Trivially, the first phase could never be carried out, since a collection "of *all* the facts would have to await the end of the world" (1966, p 11). Even all the facts up to now are almost impossible to be collected, since there is a incredible high number and variety of them. One might say that we do not need to collect *all* the facts. Rather, we need just to collect the *relevant* facts. A problem with this proposal is that we should nonetheless specify what facts are relevant to. At this point, one might rebut that facts should be relevant to a certain problem. But, as Hempel stresses, what sorts of data are reasonable to collect is not determined solely by the problem under scrutiny, but rather

by a tentative answer to the problem itself. Therefore the moral of the story is that "empirical 'facts' or findings (…) can be qualified as logically relevant or irrelevant only in reference to a given hypothesis, but not in reference to a given problem" (Hempel 1966, p 12). The idea is that data cannot be gathered without guidance provided by antecedent hypotheses. At least tentative hypotheses are needed to orient scientific investigation. Accordingly it is impossible to provide a mechanical set of rules able to generate hypotheses unless one has a bunch of hypotheses already in hand.

The second phase (b) is problematic as well. Facts can be analysed and classified in many different ways. Indeed, they should be analysed and classified according to a particular research plan. If they were not, then analysis and classification would be useless for the particular scientific investigation. For the second phase to take place, it should be based somehow on a tentative hypothesis in order to give a direction to analysis and classification. As Hempel puts it, "without (…) hypotheses, analysis and classification are blind" (Hempel 1966, 13).

Next, Hempel arrives at the core of its argument about scientific discovery. Inductive inferences are supposed to provide rules for mechanically deriving general principles from observed facts. Here the logic of discovery is identified as an inductive inference. This is the old idea of the scientific method as famously conceived by Bacon. However, as Hempel shows, a naïve conception of inductive inferences has dubious reliability. Moreover Hempel seems to imply that these rules are like algorithms. However, even in very special and local cases in which there is a procedure to derive hypotheses, the mechanical procedure for the construction of a hypothesis stems from an antecedent less specific hypothesis, which cannot obtained by the very same procedure. Actually, "scientific hypotheses and theories are not *derived* from observed facts, but *invented* in order to account for them" (Hempel 1966, p 15). In other words, hypotheses are 'happy guesses'. Moreover, Hempel supports a temporal view of the distinction, by saying that hypotheses first are freely invented and proposed, but they are accepted only after justification is provided.

2.2 Discussion: Is hypothesis generation irrational?

The problem with Popperian-like conceptions of hypothesis generation is that sometimes requirements ascribed to this phase are too demanding. Popperian-like positions overestimate *hypotheses generation* in (at least) three respects.

Hempel is representative of the first kind of overestimation. His argument against a logic of hypothesis generation stems from the idea that a logical procedure of discovery

must rely on strict logic. However, this requirement is too demanding (Nickles 1980). There is here a hidden (and unjustified) assumption that philosophical inquiries are just a matter of logical arguments, where 'logic' stands for an articulated system of rules, and it is equated with *rationality*. If you put together strict logic and rationality, then there is clearly no rationally reconstructed hypothesis generation phase. But if you stop to think for a moment in terms of algorithms and reason instead in terms of *heuristics* then there is room also for a rational phase of hypothesis generation. Heuristics should be understood as *rules of thumb* (Langley et al, 1987; Wimsatt 2007; Gross 2013) applied to find possible solutions to a complex problem. This is different from applying rules of an algorithm. An algorithm would provide just *the optimal solutions* to a problem. The notion of heuristic is related to the idea of *bounded rationality* (Bechtel and Richardson 2010), that is, the idea that in order to solve a problem the solutions we will find are limited to the initial information we have. When we start to consider hypothesis generation in terms of heuristics and bounded rationality, we become more liberal and we can start to accept reasoning modules such as analogical reasoning that can be rationally reconstructed even if they are not strict logic.

Next (the second overestimation), Hempel claims that hypothesis generation procedures should be hypothesis-free in the sense that it should be based on purely inductive organization of data. He then argues that there is no such a discovery procedure because there cannot be hypothesis-free procedures. Hempel's proposal is that data cannot be gathered without the guidance of antecedent hypotheses. But this too seems unjustified. While contemporary accounts recognize the impossibility of a completely hypothesis-free research (Leonelli 2012; Rheinberger 2011), this impracticability is not seen as a barrier to the epistemological analysis of discovery procedures. It is not clear at all why the impossibility of hypothesis-free research bans discovery procedures. For instance, if we admit that hypothesis generation depends on previous knowledge, we can nonetheless rationally reconstruct the procedure of derivation from background knowledge. Even if we are not hypothesis-free, we can nonetheless elaborate a procedure to generate hypotheses.

Finally, there is a third aspect of overestimation (Laudan 1980; Kelly 1987). Larry Laudan argues that a logic of discovery might be epistemologically relevant only if it provides epistemic warrant for the hypotheses it generates. However, this too seems to be unjustified – at least for the phase of hypothesis generation. Reichenbach's 'straight rule' of induction (1961) is a case in point as this rule might be considered as a logic for generating hypotheses. However, Reichenbach explicitly says that its inductive method should *not* be conceived as a method to produce *also* the justification of the belief in the hypothesis generated. The justification should be pursued by other means. In other words, hypotheses generated don't have to be justified. Indeed, they can turn out to be

false and this would be found out in the phase of justification.

Therefore, these considerations suggest that there can be a phase of hypothesis generation that is not irrational. It does not need to be strictly logical, but it can be based on rules of thumb that can be properly subjected to a rational reconstruction. In other words, hypothesis generation can be subjected to philosophical analysis.

2.3 Friends of discovery

The second family of positions tries to debunk the Popperian position. The position shared by friends of discovery is exactly the attitude toward the phase of discovery as *a methodology to generate hypotheses*. Some say that there is a strict logic of discovery, while others claim that hypothesis formation is more a heuristic, but the point is that for all these positions *hypothesis generation is a rational procedure that can be subjected to philosophical analysis.* Most important, friends of discovery also reveal the possibility (and somehow the necessity) for another phase in scientific discovery, thereby undermining the idea that scientific discovery is just hypothesis generation and justification. Some calls this additional phase "prior assessment" (Curd 1980), others "weak evaluation" (Schaffner 1993) or "theory pursuit" (Whitt 1990; McKinney 1995; Seselja and Strasser 2014). Despite subtle differences, the underlying idea is that once a hypothesis is generated, it has no plausibility. If we use rules of thumb to generate possible solutions to a problem, then we do not have to attach plausibility considerations. Plausibility considerations come from another source. In other words, hypotheses do not have to be justified in the very moment when they are generated. Plausibility considerations are attached to hypotheses when these are weakly evaluated.

The positions of friends of discovery were anticipated by Hanson (1975; 1958; 1960). He (indirectly) anticipates the idea of an additional phase within discovery and justification.[5] According to Hanson, hypotheses or theories might be thought in two general ways. A first way of looking at theories or hypotheses is that they are derived by typical Baconian *inductio per enumerationem simplicem, ubi non reperitur instantia contradictoria*. This was perceived as highly problematic, and Hanson recognizes that *inductio per enumerationem* wrongly suggests that laws are just a summary of data. As a matter of fact, scientists rarely find laws by enumerating or summarizing observation.

[5] Actually, Hanson's aim was to show that there can be a logic of discovery but, in failing to show this, he opened up the possibility for the phase of prior assessment/theory pursuit/weak evaluation. As shown by many critics (Nickles 1980; Schaffner 1993), Hanson is not able to demonstrate the existence of a logic of discovery in the sense of hypothesis generation, but rather he elaborates a methodology of prior assessment of existing hypotheses.

Another way is to think about hypotheses/theories as premises of a hypothetico-deductive (H-D) system. Theories and hypotheses are regarded as 'quasi-axioms' and by means of derivation some predictions will eventually follow, which are then contrasted to data. These accounts "begin with the hypothesis as given, as cooking recipes begin with the trout" (Hanson 1960, 101-2). Hanson notes that this is what usually physicists do *after* catching hypotheses. However, which are the reasons for proposing a hypothesis instead of another? According to H-D supporters there is no reason because hypotheses are generated by leaps of genius. In other words, according to Hanson both H-D and inductive frameworks are problematic, in the sense that they can hardly make sense of the processes discovery in science.

Therefore, these accounts should be rejected, or at least substantially revised because they do not explain a fundamental aspect of science: Why do we try to justify a hypothesis instead of another? Where do hypotheses come from? Clearly, hypotheses do not come out of the blue. According to Hanson one problem with these perspectives (especially the H-D account) unable to cope with hypothesis generation is that there is a conflation of reasons for *suggesting* a hypothesis into the realm of reasons for *accepting* a hypothesis. According to Hanson there is a *logical* difference between the two types of reasons. Famously Reichenbach states that any logical consideration is justificatory in nature, therefore 'suggesting' and 'accepting' are of the same logical type. Any other (non-justificatory) reason to propose a hypothesis is psychological, historical or sociological. Therefore the very issue here is whether the difference between suggesting and accepting a hypothesis is of logical type, or of degree, or psychological/sociological/historical (Hanson 1960). According to the position Hanson argues against, if one is interested in understanding the specificity of hypothesis generation procedure, then either he would consider hypothesis generation criteria as reasons for accepting a hypothesis as true (therefore implying the denial of the specificity of hypotheses generation with respect to hypothesis justification) or he would analyze psychological, sociological or historical considerations. In both cases, discovery as hypothesis generation is missed. In the first case it is conflated to justification, in the second case logical analysis is irrelevant or in general not applicable.

However, Hanson (1975, p 71) not only thinks that there is something more than psychology in hypothesis generation, but that this procedure has its own peculiar logic. Instead of relying on induction or deduction, Hanson shifts the attention to a different idea. Hanson suggests that there might be another type of inference at the origin of the discovery of theories. Following Peirce and Aristotle, he endorses the logic of *the retroductive (or abductive) inference* from a set of facts to a hypothesis (say H). The inference has the following structure:

1. A surprising phenomenon P is observed
2. P is explained only if a certain hypothesis H were true

Ergo

3. There is good reason for elaborating and going deeper in the hypothesis of the kind H

2.4 Discussion: Is it really just generation and justification?

Contra Hanson, the idea that scientific theories are discovered according to the retroductive inference is problematic. Some argue that Hanson's scheme does not show how hypotheses are generated or discovered (Nickles 1980; Schaffner 1993). It merely says *how hypotheses are evaluated in the first place*. The way hypotheses come into mind is perfectly compatible with Popper's ideas on discovery. Popper would say that the kind of inference at stake here does not generate hypotheses, because the hypothesis is not in the conclusion of the argument, but in one of its premises[6]. Achinstein (1970; 1971; 1987) and Harman (1965; 1968) also note that (a) Hanson does not consider the role of background of theories in scientific inference and (b) the fact that a hypothesis is chosen because it is more legitimate than others. For these reasons Schaffner (1993) concludes that Hanson's framework is flawed because it does not distinguish between *a logic of generation* (articulation of an hypothesis) and *a logic of preliminary evaluation* (that is what Hanson develops). In other words, Hanson shows not how a hypothesis is generated, but rather how a hypothesis is in fact weakly evaluated without being put at test as it is done in the traditional H-D framework. But if this is the case, then Hanson is not talking about reasons for *suggesting a hypothesis*, but rather of reasons for *pursuing* a hypothesis (i.e. *plausibility reasons*) because 'suggesting' implies 'generating' but here there is no logic of generation.

Some friends of discovery go deeper in this idea of 'plausibility reasons' (Nickles 1980; Curd 1980; Schaffner 1993) and they make a *distinction between hypothesis formation and hypothesis weak evaluation* showing how the phase of discovery is much more complicated than it might appear. This is, for instance, the purpose of Curd's essay (1980). According to Curd, it seems that *the discovery phase should be divided in two sub-phases*: a phase of theory/hypothesis generation and a phase of prior assessment of the theory/hypothesis generated. Therefore, the discovery/justification distinction became

[6]This is why Nickles rightly points out that "[s]ince it is not a logic of generation but takes *H* as given, Hanson's claim that retroductive inference differs from hypothetico-deductive inference is shaky" (Nickles 1980, p 23).

a *tripartite distinction*: hypothesis generation, hypothesis prior assessment, and hypothesis justification. Schaffner means the same with 'weak evaluation' (1993). Other commentators refer to prior assessment as 'pursuing' (Whitt 1990; McKinney 1995; Seselja and Strasser 2014).

The idea behind the phase of prior assessment is that theories or hypotheses are not generated in the final form to which they are subjected in the phase of justification. That is to say (as Duhem has claimed (1955, p 221)) that theories do not come out of the blue, but rather they are subjected to a 'period of incubation'. In this period of incubation hypotheses and theories are developed.

However, some raise an issue of serious concern: Why exactly should plausibility be treated differently from justification? For instance, one might accept this logical distinction but deny the conclusion that we should keep separated plausibility and justification (Salmon 1967). Consider Bayes' Theorem. As it is widely known, this theorem requires the existence of *prior probabilities* of a hypothesis. Salmon interprets reasons for plausibility *as* if they were *prior probabilities* of a theory/hypothesis. Therefore, since prior probabilities are part of the Bayes' Theorem, then the logic of justification *subsumes* plausibility. Therefore, philosophers like Salmon would say that supporters of the tripartite distinction are merely relabeling part of the context of justification as 'prior assessment'. With this trick, philosophers are pretending to talk about discovery but rather they are writing about the context of justification. An argument against this thesis might be as follows. Prior assessment and justification have different goals. Prior assessment "concerns the methodological appraisal of hypotheses after they have been generated but before they have been tested" (Curd 1980, p. 203). What kind of appraisal I am talking about? The idea is that since there is no limit to the number of hypotheses that one could derive from the body of data collected[7] then how does one decide which hypothesis to develop or to put at test? The idea of the 'prior assessment' is that some hypotheses are developed and taken seriously because they are *worth being developed* according to certain criteria (provided by the kind of prior assessment employed). Accordingly, prior assessment is necessary because without it science would be unable to choose which hypotheses, among the many derived from data, should be developed and eventually tested. A complementary argument would be that prior assessment "offers different sorts of evaluative questions than those usually considered" (1980, p 21). The cases in point are traditional theories of confirmation that consider only the acceptability of finished hypotheses. However, these theories completely ignore those epistemological considerations that lead to the development of hypotheses from being very abstract to being accurate and precise. In this phase of development some hypotheses are discarded on the basis of reasons that are different

[7] This is another way to put the idea of underdetermination (Duhem 1914; Quine 1951; Skyrms 1966)

from the one usually considered by theories of confirmation, and those reasons also specify the development of hypotheses that seem to be worth of being developed. Whitt (1990) expresses a similar idea by saying that the kind of evaluation provided by the pursuit considerations are rather distinct from justificatory reasons. This is because reasons for accepting a theory (justification) are epistemic, while reason for pursuing a theory might be also pragmatic. Claiming that a theory is promising and it is worth developing is quite different to accept a theory, and the criteria of theory choice involved appear to be pretty different. Focusing just on two monoliths (discovery and justification) may miss all the subtleties of scientific practice (Franklin 1999, p 163).

Therefore, from discussions of so-called 'friends of discovery' about Hanson's work some important concepts emerge. First, retroductive strategies are not really about hypothesis generation, but rather disclose a different kind of evaluative consideration. This different type of evaluation is what provides reasons not for accepting a hypothesis, but rather reasons for pursuing and developing it. These reasons might be pragmatic or evidential. Therefore it seems that in addition to strategies of generation and justification, there are also strategies of what have been called prior assessments/weak evaluation/theory pursuit. These strategies form an independent phase of scientific discovery. The phase is independent from hypothesis generation because during generation one does not need to put forth evidence for the plausibility of a hypothesis, while it is different from justification, because during the justificatory phase (let us call it 'strong evaluation') reasons for accepting a hypothesis are strictly epistemic.

2.5 Summary

Now we have all the ingredients for my tripartite framework. I have shown in 2.3 that hypothesis generation can be conceived as a rational procedure, if we take 'rational' as encompassing both strict logic and heuristics strategies. Hypothesis generation would be a heuristic by providing just a set of possible solutions to a problem and *not* an optimal solution. However I have also shown (2.4) that hypotheses need to be prioritized because we cannot put at test all the possible conceivable hypotheses. Moreover, I have highlighted that reasons of plausibility might be different from reasons of acceptability. Therefore, the minimal account for scientific discovery is a process involving three phases: hypotheses generation, priori assessment of hypotheses, and strong evaluation (justification) of hypotheses. All the three phases can be subjected to philosophical analysis.

3. Type of inference in hypothesis generation, prior assessment, and justification

So far, I have provided arguments in support of the thesis that, when discovering, a scientist might reasonably pass through three phases: hypothesis generation, prior assessment and justification. As Nickles notes (1980), there is no magic in number three, in the sense that one might distinguish several sub phases depending on the context of inquiry. However, a threefold distinction seems to capture a minimal core of phases that a scientist, sooner or later, will pass through. However, we should add something to this analysis. Provided that there is now a 'list' of phases delineating scientific discovery, we should start to understand what happens in each phase. The task of this section is to list the types of inference occurring within each phase. This list of inference is not intended to be exhaustive but it will be useful in the next chapters to identify the type of reasoning employed within each phase in both data-driven and 'traditional' molecular biology. I draw from the Schaffner's table of inferences in scientific discovery (1993, Chapter 2), but some points have to be corrected. Let us consider Table 1.

PHASE OF DISCOVERY	TYPE OF INFERENCE OR REASONING
Hypothesis Generation	Analogical Reasoning Derivation from Background Assumptions Leap of Genius/Intuition
Prior Assessment/Pursuit/Weak Evaluation	Eliminative Inferences (e.g. Baconian induction, eliminative induction, etc) Retroduction
Hypothesis Justification/Strong Evaluation/Hypothesis Testing	Confirmation in all its forms

Table 1. Type of inference or reasoning strategies for each phase of scientific discovery

The Popperian family is arguably right in saying that, in the phase of hypothesis generation, there are no strict logical inferences pointing to a hypothesis. In most cases, analogical reasoning plays a prominent role. Sometimes, in reasoning about a certain phenomenon, scientists derive a hypothesis about that phenomenon from an analogous case or from a completely different context. For instance Morgan and his colleagues thought about chromosomes as strings (and genes as beads on the strings) thereby facilitating the reasoning to the hypothesis "that traits with genes together on a single chromosome (…) [are] inherited together" (Craver and Darden 2013, p. 71). Lindley Darden has provided a rich discussion on the role of analogies in several works (Darden 1991; 2006). In some cases what fuels hypothesis formation is a leap of genius, or an intuition. A good example is Kekule's hypothesis of the structure of benzene. The 'ring formula' of benzene was demonstrated (justified) through the accumulation of evidence

over many years. The particular way evidence had been combined together in order to *justify* the hypothesis of the cyclic structure of benzene can be understood as the *method* of justification of the hypothesis of the structure of benzene. But the way Kekulé arrived to the hypothesis in the first place is quite peculiar. In the winter of 1861-1862 Kekulé had a dream of a snake seizing its own tail and, when he woke up, the cyclic structure of benzene came up in his mind. Sometimes it is possible to face a situation that is more prone to be rationally reconstructed. In particular, it is when a scientific hypothesis is derived as a consequence of the status of knowledge of a particular field. Traditional molecular biology is a good example of this situation. However, in most cases this derivation is done in parallel with intuition and creativity making hypothesis formation a particularly nebulous situation. It is also worth to be pointed out that in this phase, while generating several hypotheses, one does not need to provide compelling evidence for those hypotheses because, as I said, phases of discovery do not subsume justification. As suggested by Franklin (1999), how convincing the evidence for a hypothesis should be, grows gradually from the first to the last phase of justification. In other words, "[t]he evidence that might encourage a scientist to propose a hypothesis may be less convincing than that required to further pursue it, which will be, in general, far less convincing than that required for (…) justification" (1999, p. 178).

In prior assessment/weak evaluation/pursuit one is likely to find out situation that can be rationally reconstructed more precisely than those situations of hypotheses generation. There are several heuristics or inferences used to weakly evaluate hypotheses, to discard blatantly false hypotheses or simply to develop initial hypotheses. I have included also Baconian induction even though Bacon clearly thought that his method was also able to provide justificatory force, i.e. discovery subsumes justification. However, the Baconian method might be seen as a heuristic to discard false hypotheses generated in the first phase. The same applies to all varieties of eliminative inferences. Another important tool for prior assessment is retroduction. As I have shown above, though in Hanson's logic of discovery retroduction is erroneously considered as a hypothesis generation method, this type of inference provides plausible reasons to go deeper into a particular hypothesis and hence it can be considered as an inference to be pursued.

Finally, justifying hypotheses in science means providing more stringent evidence that the hypothesis under scrutiny actually holds. In other words, justifying hypotheses means looking for observational data or evidence speaking *directly* in favor of a hypothesis. This is the topic of confirmation, i.e. how we might say that a hypothesis is confirmed by evidence. The topic of confirmation is not only vast, but it is arguably the most debated issue in the philosophy of science. For the sake of this chapter I just say that there are many ways of dealing with confirmation. Putting aside famous cases such

Hempel pioneering work (1945) or Popper's falsificationism (2002), there are two big families of approaches to confirmation. The first is the hypothetico-deductive method and all its varieties (thereby including Popper's work). The second approach is inductive, and it includes all the statistical and Bayesian methods. Therefore, when treating both data-driven and traditional molecular biology phase of justification, I will look for some of the varieties of confirmation methods.

4. Evaluating and valuing

The tripartite distinction seems to capture the minimal account of how discovery is pursued. One might then distinguish, within each phase, several sub-phases, but the tripartite distinction seems to capture at least conceptually how scientists deal with scientific hypotheses. This tripartite distinction is useful because it identifies the phases of scientific discovery. Using these tools to analyze discovery strategies means looking for these phases in scientific discovery, and to understand how each of these moments is pursued. In the previous section, I listed the kind of inferences that can be used in each of the three phases. For instance, one might use analogy in hypothesis formation, while another might derive a hypothesis from a coherent body of knowledge. Prior assessment can be faced by means of an eliminative framework, while another might use retroductive inference. Finally, one might face the phase of justification by using of one the approaches to confirmation.

However, something else is missing. I have shown the abstract and conceptual temporal phases that lead a hypothesis from being formed to its acceptance. For a hypothesis to be accepted, it has first to be generated. One then might say that rarely a hypothesis is generated in a complete form and then put to the test of acceptance. Accordingly a hypothesis - once it has been generated - should be developed and refined. After that, a hypothesis can be finally "accepted," whatever this means exactly. But something else should be clarified. As I said, in the three phases hypotheses are evaluated and refined at any stage by means of a certain type of inference (e.g. induction, H-D method, abduction, etc). However, it is not still clear, whatever the inference used, the criteria establishing whether a hypothesis is a "good" scientific hypothesis. In a sense, inferences are just empty boxes, which have to be filled somehow with contents. Independently of the type of inference used, some questions must be answered: Why should scientists select a set of hypotheses instead of another? According to which criteria some hypotheses are worth developing? And in which direction should we develop hypotheses? Which are the features of a hypothesis that makes it an hypothesis "acceptable"? In other words, my questions are about criteria of theory choice.

Therefore we should note that in each phase there are criteria (which I call for now *epistemic desiderata*) that play a pivotal role in orienting research in a direction instead of another. But the tradition I will refer to, sees theory choice not as an unambiguous procedure embedded in strict rules, but rather as being embedded in a sort of value judgment. This is the topic of *epistemic values*. Although I will go deeper in the issue of epistemic values in the last chapter of this work, it is useful to introduce the topic now. The plan for this section is to introduce the notion of epistemic values and to show that, within each phase, different epistemic values can guide hypothesis generation, prior assessment and acceptability. These considerations will form one important part of the philosophical tools that inform my analyses of discovery of data-driven and traditional molecular biology in chapters 2 and 4. Simply put, in the next chapters – for both data-driven and 'traditional' molecular biology – I will see what happens in their discovery processes by tracking generation, prior assessment and acceptance *and* the epistemic desiderata orienting these phases. But, first, let us introduce the issue of epistemic values.

By borrowing a distinction from McMullin (1983), I distinguish *evaluating* and *valuing*. When we evaluate a hypothesis, we investigate whether a hypothesis satisfies certain criteria, e.g. whether it is empirically accurate or it is consistent with the corpus of certain disciplines etc. Yet, we can also evaluate a certain methodology, by analyzing if that kind of methodology is able to develop a hypothesis in a direction that would satisfy certain criteria. Unlike evaluation, when we value we do not try to understand whether a hypothesis satisfies certain criteria, but rather we discuss the criteria themselves. As already mentioned, these criteria are called *epistemic values*.

An epistemic value is an epistemic desideratum. It is 'epistemic' because it is likely to promote those characters of science that make it the type of knowledge usually seen as "the most secure knowledge available to us of the world we seek to understand" (McMullin 1983, p. 18). It is a 'desideratum' because it is something one believes will help to achieve that kind of knowledge, if adequately pursued. McMullin's iconoclastic paper (1983) aims to show that theory appraisal is a procedure much closer to value judgement than to some rule-governed type of inference. McMullin does not mean 'value judgement' in any ethical or moral sense. Although it is now uncontroversial that ethical values can play a role – even an epistemic role (Douglas 2000) – in theory choice, McMullin's thesis should be understood in a different sense. McMullin means exactly what Kuhn means by 'value judgement' in his famous *Objectivity, Value Judgment and Theory Choice* (1977) when he says *"the criteria of [theory] choice function not as rules, which determine choice, but as values which influence it"* (1977, p. 331). Therefore value judgement is to be understood not as an unambiguous procedure to determine which choice is the best, but as a *propensity* - which origin is difficult to be determined or to track down – to

consider a characteristic of an entity to be desirable for an entity of that kind. And the judgement of 'value-judgement' involves not the evaluation of how much a particular entity epitomize that value, but mainly *how much* we *value that* characteristic in *that* particular entity.

There are many epistemic values. For the moment I will draw on Kuhn's analysis, which is focused on traditional epistemic values. I will look for these traditional epistemic values within data-driven and traditional molecular biology in Chapters 2 and 3.

Kuhn (1977) makes a list of the characteristic values of a good scientific theory. He mentions five desiderata for a scientific theory that are quite uncontroversial:

1. Predictive accuracy, i.e. how accurate predictions our theory/hypothesis is
2. Internal coherence, i.e. whether a theory is hang together properly, without logical inconsistencies
3. External consistency, i.e. whether a theory is consistent with other relevant theories. A good case in point is the steady-state cosmology that violated the important principle of conservation of energy
4. Unifying power, i.e. the ability to bring together multiple areas of inquiry
5. Fertility, i.e. the ability of a theory to make novel predictions not previously part of the original agenda

Of course there can be many other desiderata[8], but the point here is that, according to both Kuhn and to McMullin, the assessment of theories/hypotheses involves value-judgement. This type of judgement is somehow essential, because value-judgement involves an appraisal of the kind of desiderata that we ascribe to a scientific theory. But one might say that, for example, whether a particular theory is fertile or it has internal coherence is not a matter of *valuing*, but rather of evaluation. However, *valuing* enters in the game because scientists "may not attach the same relative weights to different characteristic values of theory, that is they may not *value* the characteristics in the same way" (McMullin 1983, p. 16). A classic example in the history of science is the disagreement between Bohr and Einstein on the acceptability of quantum theory. Einstein negatively evaluated quantum theory because it lacked consistency and coherence with the rest of physics, while Bohr played down the importance of consistency, arguing that predictive success was more important as a criterion for theory choice. Clearly the two evaluations of quantum theory employed radically different *epistemic values* on what possibly constitutes a good theory.

While there might be criteria to evaluate whether a scientific hypothesis 'embodies' a specific desideratum, sometimes there can be disagreement on how we value a specific desideratum. *Before* we evaluate a desideratum, we should say *whether* we value a

[8] McMullin mentions also *simplicity*, but he also recognizes how problematic this desideratum might be

desideratum. Exactly this idea of valuing can provide the answer to the questions above mentioned. As a remainder these questions are:

- Why should scientists select a set of hypotheses instead of another?
- According to which criteria some hypotheses are worth to be developed? And in which direction should we develop hypotheses?
- Which are the features of a hypothesis that makes it an acceptable hypothesis?

Therefore in the tripartite distinction we should include also value judgement:

1. When we generate hypotheses by means of, e.g., background assumptions, we should motivate the desiderata leading to what we think is an optimal set of initial hypotheses
2. When we decide to develop some hypotheses, we should say which are the desiderata we value that establish which hypotheses are worth to be developed. Moreover, we should also establish the desiderata that orient in a direction or another the hypotheses themselves
3. Finally, when we accept a hypothesis as a good scientific hypothesis, we should say according to which criteria hypotheses are good scientific hypotheses

As a consequence, value-judgement is not just a matter of hypothesis acceptance, but it permeates the whole scientific process of hypothesis formation and prior assessment. Value-judgement lies at all levels: From choosing the set of background assumptions to generate hypotheses, to the choice of the experimental setting, to the final hypothesis acceptance. While the controversy motivating this dissertation is focused mainly on methodologies - as if the choice of methodologies could be decided in an unambiguous way - in this thesis I shall consider also how the structure of value-judgement informs this kind of controversies.

Therefore in the following chapters I will analyze not only the structure of the three phases of discovery in data-driven and traditional molecular biology. I will also identify the (traditional) epistemic values behind those enterprises in order to understand *whether the controversy motivating the present work is about the evaluation of epistemic values or the valuing of those values.*

5. Conclusion

In this chapter, I have drawn from the debate on discovery versus justification the philosophical tools to infer the discovery strategies of data-driven biology. I have

evaluated the debate on discovery/justification by dividing the positions in two big families, according to the attitude towards the process of discovering and developing hypotheses. The Popperian family is undermined by its tendency to look at hypotheses merely in their final form, thereby missing those scientific procedures and heuristics leading to the formation of hypotheses. The second family – the 'friends of discovery' – provides a richer picture of how scientists deal with hypotheses. In particular, they provide arguments on why we should distinguish *at least* two sub phases in the phase of hypothesis discovery. Merging virtues of both families, I have come out with a tripartite picture of how scientists deal with hypotheses: hypothesis generation, hypothesis development, and hypothesis justification. However, this picture seems too abstract, for at least one reason. Both families completely avoid explaining the criteria establishing the desiderata for a hypothesis to be worth developing, or the features that a hypothesis should have in order to be considered an acceptable hypothesis. I have then introduced the topic of epistemic values in order to say that value-judgment rather than some sort of inference establishes the criteria used to develop in a direction or to accept hypotheses. For instance one might employ a Bayesian-like framework, and saying that if a hypothesis has a certain prior probability then it is worth of being developed and if, after adding new evidence, a hypothesis reaches a certain amount of probability then it can suggest a 'justified belief'. However, what kind of evidence should we use here? In the case of Bohr versus Einstein, how do we weight the evidence of external consistency and the evidence of predictive accuracy? Analyzing these issues by means of epistemic values can make sense of the type of evidence that we value in scientific reasoning. Therefore, in the next chapters I shall analyze data-driven biology and traditional molecular biology by trying to identify in their strategies hypothesis generation, prior assessment and acceptance. I shall look for the kind of inference used in each phase. But I will complement this analysis by trying to identify the epistemic desiderata that are valued within each phase. Only by looking both at the inferential level and at the level of (epistemic) value-judgment can we accumulate enough material to compare data-driven and 'traditional' molecular biology.

CHAPTER 2

DATA-DRIVEN APPROACHES TO MOLECULAR BIOLOGY

CHAPTER ABSTRACT

In this chapter, I discuss the logic of discovery of data-driven biology. I identify the phases of the tripartite framework of Chapter 1 in typical data-driven screenings. These include genome-wide association studies and cancer genomics screenings. In particular, I highlight the role of background assumptions in hypothesis generation and the eliminative inference structure of the phase of hypothesis development. Finally, I identify the epistemic values embedded by these studies within each phase of discovery.

1. INTRODUCTION

In the first chapter, I critically reviewed the debate on the distinction between discovery and justification in order to make sense of scientific discovery and how it might be structured. The result is a set of orienteering tools useful to analyze both data-driven and traditional molecular biology. According to the philosophical tools elaborated in the previous chapter, scientific discovery is composed of three phases (hypothesis generation, prior assessment, and justification), each pursued with a specific inference (see Table 1, chapter 1) and guided by one or more epistemic values. In this chapter I analyze data-driven approaches to molecular biology by looking for the three phases and the type of inference used and by identifying the epistemic value(s) guiding each phase. In this analysis I also note that biological databases play a prominent role. Therefore in the second part of this chapter I discriminate two uses of biological databases in contemporary biology and I discuss their epistemic significance in detail.

But before starting my analysis on data-driven biology, I will recall what I have already said about data-driven molecular biology in the introduction to this work. Moreover, I shall specify also the types of case studies I consider as paradigmatic instances of data-driven molecular biology.

1.1 Data-driven molecular biology

What is a 'data-driven approach'? An approach is data-driven – as opposed to hypothesis-driven – if it designs an experiment merely to generate relevant data rather than to obtain a specific effect in order to test (*directly*) a specific causal claim. In particular, data-driven projects aim to generate specific hypotheses. Biologists like Eddy (2013) or Weinberg (2010) think that molecular biology is mainly a matter of testing claims, while data-driven supporters start from data to derive hypotheses, rather than starting from hypotheses and then going consequently to data as in the standard H-D framework. As we will see, in data-driven biology data are relevant to a corpus of background assumptions, while in hypothesis-driven biology data are relevant only to a specific and narrow claim, though the specific claim involves also other auxiliary statements.

As paradigmatic examples of data-driven biology I will employ two classes of contemporary cutting-edge studies widely used in contemporary biology: Genome-wide association studies (GWASs) and Cancer Genomics screenings. The reason to choose GWAS and cancer genomics lies in the fact that these two enterprises have become possible only through the development of technologies *post* Human Genome Project and, as a consequence, they share many of the features usually attributed to big data. Both generate astonishingly vast data sets, both are supposed to exemplify a 'hypothesis-free' way of doing science (Garraway and Lander, 2013; Vogelstein et al, 2013; Brookfield, 2010; Cortes and Brown, 2011; Guessous et al, 2011; Gorlov et al 2009; Yeo, 2011) and both make extensive use of maps as essential part of their heuristic discovery strategies.

1.1.1 Cancer Genomics

Cancer genomics is essentially the molecular study of cancer with a massive use of sequencing technologies. Several molecular studies starting from the mid 1970s showed that cancer "is the result of the activation, by modification or over-expression, of a highly-conserved family of oncogenes" (Morange 1998, p 219). The idea that genes regulating basic functions of cells when mutated might lead to cancer is known as the *oncogene paradigm*[9]. I will develop this idea in the next chapter, since it forms the core of the paradigmatic example of 'traditional' molecular biology as Weinberg means it: Molecular oncology. Therefore from the mid 1970s biologists were mainly involved in discovering cancer genes, using a piecemeal approach by trying and testing guesses about genes. Biologists are still looking for cancer genes, but the approach has slightly changed.

[9] A distinction has been developed between *oncogene* and *tumour suppressor gene*. The former is a gene that, when activated by a mutation, increases the selective growth advantage of a cell, while the latter is a gene that, when inactivated by a mutation, increases selective growth advantage of the cell

In a famous commentary, the leading scientist Dulbecco (1986) says that evidence shows that cancer is somehow linked to some "viral genes". This is the prototypical idea of 'cancer gene' or 'driver gene' that is, a gene containing mutations which confer "a selective growth advantage" (Vogelstein et al 2013, p. 1549) to the cell. However, as Dulbecco emphasizes, a piecemeal approach to discovering cancer genes would be less effective than a more comprehensive one. Having a higher system-level view of all the genes in a cancer genome would be of some help in discovering 'driver' genes. This became possible some years after the Human Genome Project, with the development of massively parallel sequencing leading to the sequencing of several exomes[10] and, in 2008, the first whole cancer genome (Ley *et al.* 2008). *Cancer genomics, by employing massive parallel sequencing, is the systematic study of the cancer genome*. Its aim is to identify genomic loci of derangement that could possibly lead the development of cancer. As Vogelstein and colleagues say, to appreciate this concept

"One must take the 30,000-foot view. A jungle might look chaotic at ground level, but aerial view shows a clear order, with all the animals gathering at the streams at certain points in the day, and all the streams converging at a river. There is order in cancer, too" (Vogelstein et al 2013, p. 1554).

Exactly as Dulbecco imagined, cancer genomics aims to have a systematic view of all mutated genes (and all the mutations) of a cancer genome. Through a systematic view, it tries to discover the driver genes and their driver mutations[11] and to elucidate the mechanisms[12] that trigger cancer development. In this sense, cancer genomics is not so different from traditional molecular studies of cancer. It has one advantage: instead of the problematic piecemeal approach emphasized by Dulbecco, it has developed the technology to obtain a comprehensive picture of the mutated genes, and hence it can discriminate more easily which genes to focus on. By 'comprehensive' picture I mean some kind of representation of *all* genes of a genome. While with the piecemeal approach we select some genes as being of potential interests for opaque reasons (e.g. a gene not investigated in the literature, an intuition, etc) with the 'comprehensive approach' we do select some genes because of statistical reasons.

As an instance of big science, cancer genomics is frequently organized in big consortia. The most famous consortia of cancer genomics are The International Cancer Genome Consortium (Hudson et al., 2010) and The Cancer Genome Atlas (Garraway and

[10] The exome is the set of the exons of a genome
[11] As Vogelstein and colleagues note > 99.9% of the alterations are simply passenger changes, and they do not increase the growth advantage of a cell
[12] Cancer genomics also deals with massive genomic rearrangements such as chromotripsis or kataegis. However, while these phenomena have been identified, it is not yet clear which are the mechanisms that can possibly explain them. See in particular the case of chromothripsis (Korbel and Campbell, 2013)

Lander 2013;). In this dissertation I shall focus on The Cancer Genome Atlas (TCGA). These projects are 'big science' projects because, taking whole genome/exome sequencing as their hallmarks, they tend to sequence as many genomes/exomes as they can. The reason for doing that is statistical. Theoretically, since driver mutations confer a growth advantage to cancer cells, they should be positively selected. If mutations are positively selected, then they should be detected more often than passenger mutations. Therefore, the more the sample size of sequenced cancer sample grows, the more it is likely to detect mutations that are statistically significant - though it increases the possibility of false positives too.

The brief history of cancer genomics has some important successful stories of mutations becoming statistically significant as soon as the sample size has been increased, and being successfully linked to genes or processes that were not at all associated to cancer. A good example is a mutation in *IDH1* (Ledford 2010). This gene is involved in cell metabolism, a process which was not associated to cancer. However, "as efforts to sequence tumour DNA expanded, the *IDH1* mutation surfaced again: in 12% of samples of a type of brain cancer called glioblastoma multiforme, then in 8% of acute myeloid leukaemia sample" (Ledford 2010, p 972). This type of discovery is possible only because data sets are big, since smaller data sets would be unable to show a robust regularity. Although scientists might be lucky and with a piecemeal approach they may discover all the cancer genes, methodologically big numbers *do make a difference*. It is also important to note that projects like TCGA might be seen also as a type of leading wedge. In the case of TCGA, this project clearly fuelled a massive development of sequencing technologies, with the consequence of price dropping for these expensive techniques[13].

To conclude this brief section, let us now give some facts about the consortium where I shall take my main case studies for cancer genomics, namely TCGA[14] (Giordano 2014). TCGA is a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The project aims to map the genomic changes in over 30 types of human cancer. By January TCGA had generated something like 1 petabyte of data about somatic mutations and structural variations from almost 10,000 cancer samples. In the next sections, I will scrutinize publications from TCGA. I will not show any particular preference for some of the laboratories involved in the TCGA, since almost all the studies of this consortium have the same structure, unveiling a strict and rigid structure of discovery.

[13] HGP had an overall cost of 2 billions dollars, while as at 2014 sequencing a genome costs few thousand dollars

[14] http://cancergenome.nih.gov/pdfs/TCGA_DataPortal_Brochure_2014

1.1.2 Genome-wide association studies

Genome-wide association studies (GWASs) fall under the category of epidemiological studies. GWASs have been defined by the National Institute of Health as "any studies of common genetic variation across the entire human genome designed to identify genetic associations with observable traits" (Manolio and Collins 2009, p 444). They scan markers across the entire genome of many individuals in order to find variants associated to a particular phenotype. In particular it is a type of epidemiological study that associates variants present in at least 1% of the population (named single-nucleotide polymorphisms, SNPs) to complex traits like common or chronic diseases (certain cancers, diabetes, hypertension, etc). The type of design of GWAS I shall consider here is the classical case-control design. Simply put, a group of individuals with a particular disease is compared to another group of individuals whose features are similar to the individuals of the first group, but they lack the disease. Individuals of the two groups are compared with respect to the presence of certain variants, namely SNPs. The goal is to see whether there is a significant difference in the allele frequency of certain SNPs within the two groups. To successfully associate some SNPs to a phenotype two conditions must be met:

1. The allele frequency of those SNPs should be above a certain threshold in the group with the disease
2. The allele frequency of those SNPs should be below in the control group

Though conceptually a GWAS is quite simple, technically speaking, it is a rather complex affair.

SNPs are scanned across *all* the regions of the genome. As of October 2014, dbSNP (the main database of SNPs) contain a total of 112 million of SNPs[15]. Hence in any GWAS epidemiologists look for all these SNPs, and then they compare the allele frequency of the two groups. Groups might be large, involving thousands of people. Therefore, it is easy to understand why these studies are instances of "big science": we have thousands of individuals examined for millions of single nucleotide variations.

GWASs are facilitated by the existence of certain big *maps* such as the HapMap Project[16]. It is the rule (as I will show also for cancer genomics) that maps facilitate data-driven screenings. A project called HapMap is of great help for GWAS. HapMap is an international project designed to provide a map of human genetic variation. As it is clearly stated in the overview of the project, the genome of any two people is 99.5% identical.

[15] http://www.ncbi.nlm.nih.gov/mailman/pipermail/dbsnp-announce/2014q4/000147.html
[16] http://www.genome.gov/10001688

However, in the remaining 0.5% there are variations that might strongly impact disease development or drug response. Among these variations, there are SNPs. Due to the principle of *linkage disequilibrium*[17], closer SNPs are inherited in block, and they are more likely to be detected together. Patterns of SNPs on a block are called *haplotypes*. HapMap delivers exactly a map of haplotype. Therefore anytime one find a tag SNP one might infer the existence of other SNPs.

The first GWAS dates back to 2005 (Haines et al 2005) although the first big GWAS was published in 2007 by the Wellcome Trust Case Control Consortium. This was a study including 14,000 patients of different common diseases plus 3,000 shared control. As of 2013, almost 2,000 GWAS have been published.

2. DATA-DRIVEN BIOLOGY

In this section I reconstruct the structure of data-driven biology according to the tripartite structure I have identified in the previous chapter. Some parts of this chapter are available also in Ratti (2015). First, I show the kind of inference used within each phase of scientific discovery of data-driven biology. I provide a great deal of detail about the discovery strategies of both GWASs and cancer genomics. Next, I identify epistemic values playing a key role within each phase of discovery.

2.1 The tripartite framework embedded in data-driven biology

In my interpretation of the practices of contemporary biology, data-driven molecular biology fits particularly well in the tripartite framework proposed in the first chapter. More specifically, data-driven biology is composed of three discernible phases:

a) Formulation of a set of competing hypotheses
b) Elimination of false (or less probable) hypotheses, and development of more probable hypotheses
c) Test (validation) of hypotheses not eliminated in phase (b)

(a) and (b) bear resemblances to the 'eliminative inductive' framework (Norton 1995; Earman 1992; Kitcher 1993). My claim is that the first two phases of data-driven biology (namely hypothesis generation and prior assessment) are guided by a special kind of

[17] Linkage disequilibrium is the "is the nonrandom association between alleles at different loci. (…) Generally, loci that are physically close together exhibit stronger LD than loci that are farther apart on a chromosome" (Visscher et al 2012, p. 8).

eliminative induction called 'eliminative inference'. This might sound strange, since eliminative induction is taken to provide also the necessary justificatory weight, thereby spanning the three phases. It has been argued that sometime eliminative inferences cannot properly establish the acceptability of a hypothesis (Forber 2011). Nonetheless, eliminative inferences provide a sort of evaluative criterion, though weaker than theory choice. Therefore, it seems that eliminative inferences might play also the role of theory pursuit, and not always the role of theory choice. However this argument should be discussed a little bit more. Hence before entering into the detail of data-driven practices, I would like to spend a few words on eliminative induction.

2.1.1 Eliminative inferences

Eliminative induction is a strange creature in the history of philosophy. In contemporary debates, it is also known as 'induction by means of deduction' (Hawthorne 1993), 'eliminative inference' (Forber 2011), or 'strong inference' (Platt 1964). An early example of eliminative inferences might be found in Bacon and his methodology, based on the comparison of the table of presences, absences and degrees. Another famous example is the set of methods elaborated by Mill such as the method of agreement, the method of difference, the joint method of agreement and difference, the method of residues and the method of concomitant variation. All these methods start with a list of candidates for potential causes producing a certain phenomenon, and they all proceed by eliminating some of them. Finally, whatever remains is taken to be the true cause. Let us see first a standard picture of eliminative inferences.

One builds a space of hypotheses/theories/possibilities about a certain state of affairs by means of a set of background assumptions or premises. Then one uses observation/evidence to eliminate all the hypotheses in that space but one. It is the Holmesian way to truth namely that "when you have eliminated the impossible, whatever remains, however improbable, must be the truth". As already emphasized, in the standard picture the eliminative step provides not only the evidence for the falsity of some hypotheses, but also the necessary justificatory weight for the remained hypothesis *to be accepted*. In other words, eliminative inferences guide *theory choice*. Depending on the author, each phase is carried out in a particular way. Earman (1992) and Norton (1995) think that the space of possibilities is built by taking into consideration the local features of the community of practising scientists, while according to Kitcher (1993) the set of hypotheses depends on the practices of the individual scientist. Earman (1992) and Norton (1995) accept both inductive and deductive procedures in the elimination step

while Kitcher (1993) opts in favour of a purely deductive elimination. However, some issues challenge the standard picture.

First, it is not entirely clear in which sense eliminative induction is "inductive." To be inductive, sometimes it is said that an inference should be ampliative. However, it is not clear how eliminative induction is ampliative. One might say that the conclusion of eliminative induction goes beyond observation, because assumptions *plus* observation entails the conclusion. However, it is possible to put the eliminative argument in a pure deductive form, with assumptions forming one premise and observation forming the other one. I prefer to use the label 'eliminative inference' (Forber 2011) instead of eliminative induction because it is more neutral with respect to the supposed inductive nature of eliminative induction.

A second concern is the problem of *unconceived alternatives* (Stanford 2006). To put it simply, if serious hypotheses "are left out of the constructed possibility space, then eliminative inferences may produce misleading or mistaken theory choice" (Forber 2011, p 189). There is no clear solution for the issue of unconceived alternatives but as I will show, in some cases data-driven biology can avoid this problem. Related to this issue, there is the challenge that some tasks have an infinite number of hypotheses, such as determining the exact value of a physical constant. Forber proposes a pragmatic solution to this by claiming that "the infinite number problem can be overcome by imposing a finite partition on possibility space" (2011, pp 188-189).

On the side of the eliminative step, there are problems too. The first is *testing holism* (Duhem 1954). The idea is that no hypothesis alone can make contact with observations, in the sense that one needs a whole body background theory to confront theory with data. Since the network of statements embedded in a theory is vast, any conflict between theory and data might be reconciled by changing some elements in the network. However, this is not just a problem of eliminative inferences, but of any assessment of evidence and, to date, it has no clear solution.

Another problem – in my opinion the most important for the present work – is that evidence in science rarely produces *tout court* elimination. Often, evidence is of a probabilistic nature. Forber (2011) says that perhaps eliminative inferences do not directly establish theory choice, but rather they establish the boundaries for such a choice. In Holmesian terms, we might say that eliminative inferences do not identify the murderer, but rather a list of highly potential *suspects* for the murder. It also establishes that, in the initial list of suspects, some of them are likely to be innocent.

This consideration fits very well with the tripartite framework of the first chapter. By various means we generate a universe of hypotheses and with the help of eliminative inferences we eliminate hypotheses that are very likely to be false, and we end up with a final set of hypotheses, that have to be strongly evaluated or corroborated by other

means. Therefore, if eliminative inference frames theory choice by providing a set of good candidates, and if theory pursuit/prior assessment/weak evaluation establishes which hypotheses are worth developing, then eliminative inferences are a type of theory pursuit/prior assessment/weak evaluation. Forber says that elimination becomes then part of the construction of the possibility space, rather than evaluation of evidence. In a sense, this is compatible with my perspective. Since for Forber evaluation of evidence is part of the justificatory phase, elimination is not part of this context, but of another one: The context of discovery, where we generate a final and complete set of hypotheses to be strongly evaluated.

In what follows, I shall apply to data-driven biology a version of eliminative inferences that is compatible with Forber's perspective, namely that it is better to frame eliminative inferences as a process of *prioritization* of theories and hypotheses rather than theory choice. 'Prioritization' implies exactly the idea of an additional procedure aimed at providing more stringent evidence (or evidence of another kind) for what has been prioritized. Therefore, one might frame eliminative inferences embedded in the tripartite framework of the first chapter phases as follows:

(a) The generation of a preliminary set of hypotheses from an established set of premises
(b) The prioritization of some hypotheses and the elimination of others by means of other premises and new evidence
(c) The search for more stringent evidence for prioritized hypotheses

Let us now start to situate data-driven biology into this framework. In data-driven biology, hypotheses are usually conjectures about entities or activities that might causally contribute to the production of a phenomenon. The idea is that each entity can be thought of as being one cause (among many) that can contribute to the development and maintenance of a biological system. However, most of the phenomena investigated are produced by the interplay of several entities. Therefore in the initial universe of hypotheses there will be more than one true hypothesis. Premises take the form of 'background assumptions' in providing valuable guidelines to build the initial set of hypotheses. Hypotheses prioritized in phase (b) should be validated in phase (c), in the sense that the way entities causally contribute to the phenomenon of interest should be clearly identified. The causal role of entities into the phenomenon of interest is framed in terms of the contribution of entities to the production of the phenomenon of interest.

2.1.2 Hypothesis generation

Let us start with phase (a), namely hypotheses generation. In this phase background assumptions (or, as Norton calls them, 'set of premises') identify a set (a universe) of competing hypotheses extracted from a data set. Hypotheses are about the causal contribution of entities to the production or maintenance of phenomena.

The first point to note is that phase (a) is not a hypothesis-free step. The impossibility of a completely 'hypothesis-free' scientific research is vastly acknowledged in the philosophy of science literature both in recent and less recent times (Hempel 1966; Popper 2002; Rheinberger 2011; Leonelli 2012b; Chapter 1 above). The general idea is that data cannot be gathered without the guidance of antecedent hypotheses because one would have no basis to identify relevant data without *at least* preliminary guidelines. Therefore data-driven research must make use of hypotheses of various kinds. This minimal requirement is inescapable. I will call these 'tentative hypotheses' *background assumptions*.

Indeed, any data-driven research has certain background assumptions. However, these background assumptions play a weaker role than hypotheses or theories play in, for example, the standard hypothetico-deductive (H-D) method. In order to explain what I mean, I introduce the distinction (suggested in the context of the literature on exploratory experiments) between *theory-driven* and *theory-informed*. According to Waters (2007), an experiment is theory-driven (or hypothesis-driven) when theories/hypotheses influence the experimental design in order to answer to a specific question, like the test of theories or hypotheses themselves. Differently, an experiment is theory-informed when theories or hypotheses do not provide *specific* expectations or anticipations of the results that will be achieved, and experimental designs are not set up in order to generate a specific effect. 'Theory-informed' experiments are used not to test a pre-existing theory. Rather, the role of theories/hypotheses is to provide guidelines and suggest strategies aimed at generating significant findings about a phenomenon when a pre-defined theory is absent. Phase (a) in data-driven biology is *theory-informed but not theory-driven*. Background assumptions, by providing loose guidelines, define which data are relevant and which are not. In data-driven biology, background assumptions are not qualified to 'test' themselves (as experiments in H-D are designed to test the initial hypothesis), but rather data-driven biology *makes use of the assumptions* without questioning them. By providing guidelines to select relevant data, background assumptions also specify the kind of hypotheses that will compose the universe of hypotheses to be narrowed by eliminative inferences. Let us see how.

GWASs make use of important background assumptions (Kitsios and Zintzaras 2009). The first and foremost important assumption of a GWAS is the 'common disease

common variant' assumption. Simply stated, the assumption holds that common variants likely to cause a certain disease are to be found in most of the human populations that manifest the disease. This assumption is related to another one: Single nucleotide polymorphisms (SNPs) are responsible genetic variants of diseases (or, at least, proxies to causal variants). As previously explained, SNPs are single nucleotide variants whose alleles are present in at least 1% of the population. In other words these two assumptions suggest precisely which facts a GWAS should be interested in in the first instance. The first assumption says that genetic variants are implicated somehow in a disease, and we should investigate them. The second assumption says that, among the types of variations, one should be interested in SNPs, and *not* in other types of variants. These two assumptions make GWASs *theory-informed*, in the sense that they supply an important background to select relevant data. However, by distinguishing relevant from non-relevant data, these background assumptions also draw the boundaries of a (finite) universe of competing hypotheses: All the SNPs might be, potentially, responsible for the disease. This means that researchers hypothesize that all the initial detected SNPs can be causal variants. In other words by using a SNPs array the initial universe of hypotheses of a GWAS is composed by millions of hypotheses (i.e. by millions of SNPs).

The problem of unconceived alternatives here is an issue. As a matter of fact, SNPs are identified in a screening because there are other studies (such as HapMap) identifying a certain variant as a SNP. The problem is that once we do a screening we cannot know *a priori* whether we have detected all the SNPs, and hence whether we are missing some important variants that have an effect on the phenotype of interest. On the other hand, the problem of the infinite alternatives *is not a concern*: However high the number of SNPs detected might be, there is nonetheless a finite number of SNPs.

Cancer genomics has background assumptions too. The first assumption is that cancer is a phenotype driven by mutations that accumulate in the genome through the entire life of an individual. Therefore, cancer genomics is interested in somatic (acquired) mutations, and not in germline mutations[18]. However cancer genomics is interested only in a subset of somatic mutations. These are the so-called *driver* mutations, i.e. mutations that drive cancer development in the first place, by providing selective advantages to the cells carrying them. This implies that cancer genomics will look for driver mutations within the set of relevant facts, i.e. somatic mutations. Although the interest in somatic mutations is clearly theory-informed, this assumption says nothing about which somatic mutations in a cancer sample will be driver or not. As in the case of GWASs, the background assumptions merely specify the space of hypothetical solutions to the

[18] This is an assumption of consortia (e.g. The Cancer Genome Atlas) in cancer genomics. However, other studies of molecular oncology might be interested in inherited mutations (e.g. studies in the famous heritable retinoblastoma), and not the ones acquired through individual's life

particular scientific problem at hand, but they do not say which of the possible solutions is the right one. The initial (somatic) mutations detected, as in the case of SNPs in GWASs, represent the universe of hypotheses that will be narrowed by eliminative inferences.

In the case of cancer genomics the issue of infinite alternatives is not harmful. As in the case of GWASs, however big is the set of somatic mutations, this would be nonetheless finite. Concerning the problem of unconceived alternatives the situation is more complicated. In principle, since we compare cancer samples either with a normal sample or with the Reference Genome, we should be able to detect all the mutations. However, in practice this might not happen. It can be the case that, for instance, we do not know that in the normal sample there is a somatic mutation that alone is not able to trigger cancer development, but that it has some effects in the phenotype. In this case, by comparing it to cancer sample (assuming that in the cancer sample there is the same mutation), we would miss that mutation because it would not count at all as a mutation.

A final comment is in order. What kind of inference is used in the hypothesis generation phase of data-driven biology? Unlike other biological disciplines, in data-driven biology the role of both analogical reasoning and intuition is particularly limited. In data-driven biology hypotheses are neither generated by means of drawing analogies from neighbour fields nor there is place for intuitions or leaps of genius. Given the highly regimented nature of the construction of the initial universe of hypotheses, in data-driven biology we might say that hypotheses are somehow derived mostly by background assumptions. As I have shown, in GWASs or cancer genomics we generate hypotheses that are consistent with a corpus of knowledge about the nature of SNPs or somatic mutations. The role of background assumptions is then complemented with actual as observation of samples, in the sense that we generate hypotheses about variations if and only if we detect those variations.

2.1.3 Eliminating and developing hypotheses

In the second phase, eliminative inferences are used to narrow the finite universe of hypotheses (i.e. to *eliminate* false or less probable hypotheses). Hypotheses that are not eliminated by eliminative inferences are then developed.

2.1.3.1 Eliminating hypotheses

In the previous section I explained that background assumptions in GWASs establish an initial universe of entities that may be responsible for a particular phenotype (the initial millions of SNPs). However, any epidemiologist knows that most of the SNPs cannot be

responsible for the phenotype. Hence, practitioners make use of a set of criteria and observations to eliminate SNPs that are not causal variants. Let us see which criteria are used. In a typical GWAS a group of individuals with the phenotype of interest is compared to another sample of individuals. The individuals of the second group are similar to the individuals of the first group, but they lack the phenotype of interest (e.g. diabetes). The core of the procedure is to see whether there is a significant difference between the two groups in the allele frequency for each SNP. When I say 'significant' I mean that it has to be higher than a particular threshold, called 'significance level[19]'. If the proportion between the frequencies in the two groups of a particular allele of a SNP exceeds the significance level in favour of the group with the phenotype of interest, then the variation is taken to be associated to the disease. Thus, elimination of SNPs is done by means of the 'significance level' *plus* the actual observation of the two groups of samples.

I should stress an important point about elimination in GWASs. As I said, a SNP is correctly associated with the disease if the proportion of the frequency of a variation exceeds a significance level *x*. If a SNP (a hypothesis of the universe of hypotheses) does not do that, then it is discarded. But GWASs are local, meaning that results are very sensitive as to how the experiment is designed. In particular, sample size tremendously affects results. For instance, a SNP might not be associated to a phenotype *x* in a study with a sample size *z*, but it can be associated to *x* in a study with a bigger sample size. This means that the argument for eliminating entities from a universe of hypotheses is contextual to the particular data set considered. Therefore SNPs are discarded only in a probabilistic sense: If the proportion of a variation of a SNP does not exceed the significance level, then it has a lower probability of being causally relevant to the disease, and it is discarded from the universe of hypotheses. But the study, due to experimental design, *might be wrong*. This is why SNPs are not eliminated *tout court*. Rather they are not *prioritized* in a particular study with a particular data set. This is one of the limitations of eliminative inferences, namely that evidence is always statistical and it cannot provide sufficient evidence for accepting a hypothesis as well sufficient evidence to discard one. On the side of accepting a hypothesis, much more is required.

This preliminary statistical analysis could not be sufficient to eliminate spurious associations. Therefore, the first statistical analysis is complemented with other two kinds of statistical analyses. First, there is the so-called 'technical derivation'. This is the re-analysis of a GWAS samples using another platform. This, in principle, should purify the previous eliminative phase from technical errors that might lead to spurious associations. After this step, GWASs come out with a universe of selected hypotheses that has to be narrowed promptly in the so-called 'replication' (Hunter et al. 2008). In the replication

[19]How to choose the significance level is a matter of debate and usually it varies according to the particular experimental design employed, sample size, etc

step, SNPs that have been so far resistant to the eliminative 'axe' are tested in additional samples to see whether SNPs are associated to the disease in the same way as shown above. This is an attempt to provide a stronger evaluation of hypotheses, but still not good enough for biology, as we will see.

Recently GWASs have started to make use of other important 'eliminative principles' (Schaub et al. 2012; Boyle et al. 2012). These principles are *biologically driven*, in the sense that they are derived somehow from the body of knowledge in the field. The problem with GWAS results is that, while some SNPs fall within coding regions (and so their precise function can be hypothesized according to the genes that they target), many others lie in(?) non-coding regions. The functions of non-coding regions with respect to gene regulation are particularly challenging to determine. However, a 'map' has recently provided some hints. The ENCODE project (ENCODE 2012) has provided annotations for all the biochemical activities within the human genome almost at a nucleotide resolution. This means that it is possible to see whether SNPs that were not eliminated (that is, that are prioritized) in the previous phases, are situated in a non-coding region that "overlap a functional region or are in strong linkage disequilibrium with a SNP overlapping a functional region" (Schaub et al. 1749). If a SNP falls in a region of the genome, and the (biochemical) activity of that region has nothing to do with the phenotype investigated, then it can be eliminated from the universe of hypotheses. In other words,

"ENCODE (…) does not only say 'these are the parts to be considered', but proposes, for each, very specific hypotheses to be investigated" (Germain et al. 2014, p 14).

Hence many SNPs are eliminated from the universe of hypotheses either because they fall within a non-functional region of the genome, or because they are located in a region which function has nothing to do with the phenotype of interest.

With all these procedures, variants that are spuriously associated to the phenotype of interest are eliminated, and the remaining SNPs correspond to the final universe of entities very likely to be responsible for the phenotype of interest.

Similar procedures may be drawn for cancer genomics as well. In this field, the processes through which the set of somatic mutations is filtered out are called 'prioritizations' (Raphael *et al.* 2014). But how does a scientist choose candidate driver mutations? There are several methodologies employed (Raphael et al. 2014; An et al. 2014). As an example, here I focus only on one. As mutational processes converge to a common oncogenic phenotype, "the mutations that drive cancer progression should appear more frequently than expected by chance across patient samples" (Raphael et al 2014, p 7). The reason is that, since driver mutations confer a growth advantage, they are positively selected. However, it is necessary to define what 'more frequently' exactly

means. This is why, in each high throughput screening, a background mutation rate (BMR) is calculated. BMR provides an example of eliminative principle for cancer genomics. As in the case of the 'significance level' idea, BMR guides elimination of somatic mutations with the help of actual observation, by means of comparison of cancer samples with either normal sample or the Reference Genome. Also here, observation is clearly theory laden. The idea behind BMR is that a mutation, in order to be a candidate driver mutation, should be present at a rate that is higher than BMR. If the mutation does not do that, then it is discarded. This means that the universe of somatic mutations is narrowed in the first instance by eliminating all those mutations that are below BMR. However, also cancer genomics is local, and its results depend on sample size. This means that somatic mutations above BMR are *prioritized*, while somatic mutations that are below BMR have less probability of being drivers. However, a study with a bigger sample size (and hence a different BMR) might show that a mutation discarded in one study should be prioritized[20]. Again here, the probabilistic nature of eliminative inferences arises.

The principle of BMR is complemented with the idea that candidate driver mutations are likely to target genes that are mutated at a rate higher than a BMR designed specifically for genes. In addition to BMR, when genes are concerned there is an eliminative principle that is biologically driven. In order to eliminate mutated genes (and as a consequence other mutations) a typical standard is to check whether genes mutated at a sufficiently high frequency significantly overlap with known cancer pathways (Vandin et al. 2012). Therefore one may say that, if a candidate driver gene does not overlap with a known gene pathway, then it is discarded. For example, Lawrence and colleagues (2013) eliminate several recurrent genes (and as a consequence also mutations) from the universe of initial hypotheses because these do not participate of any known cancer pathways (they have functions that, so far, are supposed to have nothing to do with cancer). Again, this is imperfect and only probabilistic. Nonetheless, with similar procedures, cancer genomics provides a final list of genes and mutations that are candidates for being drivers, and they are then validated experimentally in the next phase.

2.1.3.2 Developing hypotheses

Both in cancer genomics and GWASs, there is not just the elimination of less probable hypotheses. More probable hypotheses are developed as well. This is done especially in the second part of the eliminative process, i.e. the biologically driven process. In phase

[20] There are also cases of driver mutations that are not recurrently present, but it is rare

(a), hypotheses have the abstract form such as 'the SNP *x* is associated to the phenotype *y*' or 'the somatic mutation *x* is associated to cancer *y*'. However, such an abstract type of hypothesis is of little use to biologists. These empty hypotheses should be filled with biological contents in order to be properly prioritized. This is done by means of comparison with biological databases or maps such as the ENCODE Project.

One strategy to fill with contents abstract hypotheses is to speculate that a certain variant or somatic mutation influences gene expression or the exonic sequence of the gene itself. Then one identifies the function of the gene mutated or close to a certain variant, and proposes a specific hypothesis of how that particular function might influence the phenotype of interest. Therefore the initial abstract hypothesis is complemented with biological information that can provide clues on how to set up experiments to be done in phase (c). The use of databases is pervasive in this type of hypothesis development. Consider for instance GWASs. First, the fact that one can locate SNPs along the genome depends on maps developed by projects such as the Human Genome Project. Without this epistemic contribution, one would not be able to develop a hypothesis about which genes a SNP might target. Moreover, only with these 'biological maps' it is possible to say whether there are genes *at all* in a region of the genome, and to have a sort of 'ID'. Finally you identify the function of the closest genes by consulting a database (although you might run an experiment). Without this information, one would not be able to hypothesize whether the gene targeted by a SNP might have something to do with the phenotype of interest. Therefore with the comparison of SNPs with maps and databases, one might develop the hypothesis 'the SNP *x* has an effect on the phenotype *y*' to the more precise hypothesis 'the SNP *x* regulates (for instance) the overexpression of the gene *MYC* which in turn with the process *z* has an effect on the phenotype *y*'. The latter hypothesis clearly facilitates the design of a more precise experiment. Consider for instance Figure 2.1. This is a screen shot of the location of the SNP rs522444. With this kind of representation, information stored in biological databases might be of some help in reasoning on which genes rs522444 may possibly influence.

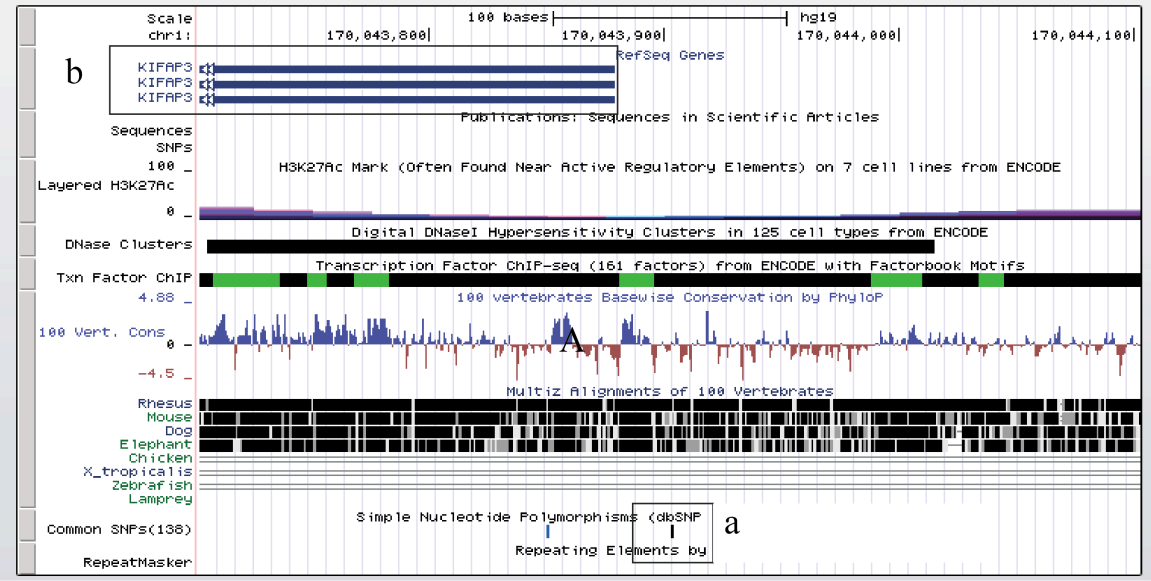

Figure 2.1. The map of the genomic region where a SNP such as rs522444 may lie.(a) the location of the SNP rs522444 (b) KIFAP3 is the closest gene to the SNP rs522444. This 'map' provides clues on what kind of functional impact (if any) the SNP might have and which gene (if any) it might target. Without this 'map' making these hypotheses would be difficult.

2.1.3.3 The phase of hypothesis-testing

At the end of phase (b), we finally have well-developed hypotheses. The biological entities involved in these hypotheses are taken to play a causal role in the phenotype of interest. In phase (c), these causal roles are 'strongly' evaluated, tested, and validated. This is the 'hypothesis-driven' phase. Traditionally, in molecular biology scientists strongly evaluate a hypothesis of the kind, "the entity *x* has a causal role in the production of the phenomenon *y*" by discovering the mechanisms of production of *y* and the role of *x* in them. There are many rational paths that lead to the discovery of mechanisms (Bechtel and Richardson 2010; Craver and Darden 2013). I shall examine the mechanistic philosophy in the next chapter, but it should be noted again that Weinberg and other 'traditional' molecular biologists clearly refer to these methodologies when they refer to the discovery strategies of 'traditional' molecular biology.

The gold standard for hypothesis acceptance in molecular biology is a mechanistic description of how *z* is implicated in the phenomenon (with exceptions, see Boniolo 2013). In order to derive such descriptions, there are several strategies. Strategies particularly close to the gold standard of molecular biology are, trivially, the *experimental approaches*. Craver and Darden (2013, especially chapter 8) present the general strategies used. These involve intervening on a specific component of a system. By observing the consequences of the intervention on an entity, its contribution to the whole system might be inferred. I argue that in data-driven molecular biology practitioners

move within the conceptual horizon of mechanistic descriptions (they sketch rough mechanistic explanations) and they use experimental intervention on a component to observe consequences. However, unlike traditional molecular biology who seek to understand also the mechanistic tiny details of the consequences of the intervention, data-driven biologists usually tend to collect just few experiments of this sort without going into much detail. In other words, data-driven biologists in the justificatory phase make use of particularly weak (and naïve) confirmation frameworks.

In GWASs, the work of most screenings usually finishes with phase (b) (see for instance Ripke et al 2013). However, practitioners know very well that "the confirmed signals emerging from GWAS scans and subsequent replication efforts are just that – association signals. The causal variants will only occasionally be among those" (McCarthy et al 2008, 365). This is why there is the urgent need "to obtain functional confirmation that the variants implicated are truly causal, and to reconstruct the molecular and physiological mechanisms" (McCarthy et al 2008, 366). There are studies (e.g. Pomerantz et al, 2009; 2010) that select SNPs associated to a disease in many GWASs and try to fully achieve (c). Freedman and colleagues (2011) provide a review of the strategies used (and strategies that should be used) in the so-called 'post-GWAS' phase (that is, phase (c)). Prominent among the strategies is the experimental manipulation of selected model organisms in order to implicate transcription units to the associated loci found in GWASs. This kind of strategy is the one used by data-driven biologists to fulfil phase (c) or, at least, to approximate its achievement. For instance Pomerantz and colleagues (2009) seek to understand the causal role of the variant rs6983267 in colorectal cancer pathogenesis. This variant has been associated with colorectal cancer by several GWASs. Pomerantz and colleagues are not simply driven by background assumptions: They are really 'hypothesis-driven' because they devise experiments in order to obtain a specific effect which in turn should provide evidence for the specific hypothesis 'the 8q24 cancer risk variant rs6983267 has a causal role in colorectal cancer by deregulating its closest gene *MYC*'. Another article, again by Pomerantz and collaborators, analyzes the variant rs10993994 associated with prostate cancer risk by several GWASs. By *in vitro* analyses, Pomerantz and colleagues find out that rs10993994 actually influences the expression of two genes thereby increasing the risk of prostate cancer. Again, experiments have been devised in order to test the specific hypothesis 'the variant rs10993994 has a causal role in prostate cancer by influencing its nearest genes'. Although these studies move with the conceptual horizon of mechanisms, confirmation of hypotheses rely on far fewer mechanistic details than in traditional molecular biology.

These considerations are more evident in cancer genomics. Also for cancer genomics, especially in the work of TCGA, screenings end with (b). If you consider the

growing list of publications of the TCGA[21], most of the screenings are exactly aimed at providing comprehensive characterizations of various tumours and to generate a list of recurrent mutations and candidate driver genes. However, there is also a tendency to 'close a story', in the sense that one generates (by means of exome or whole genome sequencing, or even with microarrays) a list of candidates, and finally she picks up one to be strongly evaluated (validation). For instance Zang and colleagues (2012), after sequencing 15 exomes of gastric cancer samples, come out with few candidate driver genes and, in particular for *FAT4*, they provide several experimental evidence that the gene is a cancer gene. For example, by silencing *FAT4* in selected cell lines they observe a significant cell proliferation (a traditional hallmark of cancer), and an enhanced cellular invasion and migration (a distinctive feature of metastases). This line of evidence does not provide careful mechanistic descriptions, but at least it strongly suggests that the mutations of *FAT4* may be drivers. Liu and colleagues (2014) analyse several gastric tumours, and they identify a particularly upregulated gene isoform of *ZAK*, that is then 'validated' as *FAT4* is. Another example is (Davis et al 2014), a study on chromophobe renal cells carcinoma that leads to the identification of *TERT* and its validation. Others use the data of TCGA to do phase (a) and (b) and then come out with (c). An example is the overexpression of *MTBP* in breast cancer detected in TCGA data sets and subsequent *in vivo* analyses of this candidate (Grieb et al 2014). All these 'validations' are not aimed at providing detailed mechanistic descriptions, but they move within the same conceptual horizon.

2.2 Epistemic values in data-driven biology

Now it is time to explicitly list the epistemic values influencing each phase of discovery in data-driven molecular biology. Epistemic values are those tendencies influencing theory choice – as well as hypothesis generation and prior assessment. To be sure, epistemic values do more than just influence theory choice: they basically establish those criteria that fill scientific inferences, which are actually empty and somehow formal. As I asked in the first chapter: Why should scientists select a set of hypotheses instead of another? According to which criteria are some hypotheses worth developing? And in which direction should we develop hypotheses? Which are the features of a hypothesis that makes it an acceptable hypothesis? These are questions that can be answered by referring to epistemic values, known also as cognitive values, constitutive values or virtues of a theory.

[21] http://cancergenome.nih.gov/publications/TCGANetworkPublications

Epistemic values, by definition, are more a matter of valuing than evaluating (McMullin 1983). We might establish quite precisely whether a particular hypothesis or theory satisfies a theoretical virtue such as coherence, but it is a completely different question whether an epistemic value is valuable *per se*. In rest of this section, I shall explain which epistemic desiderata are valued within each phase of data-driven biology.

2.2.1 Epistemic values playing a role in all levels

As with 'traditional' molecular biology, in data-driven biology the epistemic value of 'internal consistency' is ubiquitous at all levels of inquiry. Internal consistency is a sort of prerequisite for any rational endeavour. In other words, internal consistency is a virtue that any scientific theory should necessarily embed. If a scientific theory or hypothesis is not consistent with itself, then anything can follow for the very simple principle of *ex contradictione sequitur quodlibet*[22]. Therefore, internal consistency is assumed at all levels of the structure of discovery (hypothesis generation, prior assessment, justification).

Another epistemic value playing a prominent role at all levels of inquiry both in traditional molecular biology and data-driven biology is *empirical adequacy*. The fact that the statements derived from a hypothesis must be consistent with observation is a prerequisite for any empirical science, and *biology is an empirical science*. These ideas might be found also in other traditional loci of philosophy of science, such as (Hertz 1899).

2.2.2 Epistemic values in hypothesis generation, weak evaluation and validation

There are three key epistemic values influencing hypothesis generation.

First, there is external consistency. As I have shown, in GWASs the fact that the initial detected SNPs constitute the initial set of hypotheses depends strictly on the background assumptions. I noted that hypothesis generation is theory-informed, in the sense that a theory forms the background and the space of possibilities where hypothesis generation moves. This means that the initial selected set of hypotheses depends on the existence of a certain corpus of notions, and that this initial set must be consistent with that corpus. If there were no corpus of notions, there would be no hypothesis generation. Hypothesis generation then should be consistent with some prerequisites in the form of a background theory (theory loosely conceived). In the case of GWASs, the set of initial hypothesis should be composed of SNPs, since background assumptions hold that to solve

[22] Assuming that we presuppose classical logics

the issue of finding the genetic basis of common diseases, we should look at SNPs. Therefore, the initial universe of hypothesis should not be formed by other types of variations. The same applies for cancer genomics. The set of hypotheses we initially select (being either mutations or genes) depends on the fact that mutations and genes are taken to be the entities to look for in molecular oncology. Therefore, the initial set of hypothesis should be consistent with the idea that mutations (and mutated genes) drive tumour progression.

The other two values in hypothesis generation are, as I have already anticipated, internal consistency and empirical adequacy. While internal consistency is pretty self-explanatory, empirical adequacy should be explained a little bit in this context. We cannot select a SNP from a database of SNPs (e.g. dbSNP) as part of the initial set of hypothesis if that SNP is not detected in the screening. If we do this, we would derive the consequence that this SNP might play a role in the development of the disease under scrutiny but this will stand at odds with the observable fact that *there the SNP is not in the data set*. The same applies for cancer genomics. We cannot include a gene in the list of genes that possibly plays a role in the cancer we are studying if that gene is not mutated in my cohort. Again, we would derive the unpleasant consequence that that gene might be driver in contrast to the observation that the gene we are considering *is not mutated at all*.

In weak evaluation, internal consistency again is a prerequisite, as well as empirical adequacy. Concerning empirical adequacy, in a GWAS we cannot continue to pursue the hypothesis that a SNP plays a role in disease development if that SNP allele frequency is below the statistical threshold of a GWAS. This would have the consequence of claiming that this SNP, while being spuriously associated to the disease, actually has features that cannot be ascribed to spuriously associated SNPs. The same applies for BMR and mutations/genes in cancer genomics.

Most important, external consistency plays a prominent role also in the phase of hypothesis development. Actually, the very possibility of the development of a hypothesis of the form "the SNP x plays a role in the disease y" or "the mutation x plays a role in the development of the tumour y" stems from an operation rooted in external consistency. The comparison of certain bits of data with functional annotations of biological databases is exactly a matter of external consistency. The idea is that we should develop the abstract hypothesis "the SNP x plays a role in the disease y" by providing suggestions on how the SNP actually does what it does. We do this by suggesting that the SNP can have a certain function. Annotations on databases such as RegulomeDB are clearly of some help: We fill the black boxes in abstract hypotheses by borrowing information from similar databases. Therefore the way hypotheses are developed is first of all consistent with the corpus of biological knowledge stored in biological databases, *because the corpus of*

knowledge represented by biological databases is exactly where useful information is extracted in the first place. We face a similar situation in cancer genomics where the recurrently mutated genes should overlap already known cancer pathways in order to suggest in what sense a gene might be driver.

In hypothesis validation, internal consistency is (again) a landmark in this phase. However, the most important epistemic value in this phase is clearly empirical adequacy. The fact that complete hypothesis of the form 'the 8q24 cancer risk variant rs6983267 has a causal role in colorectal cancer by deregulating *MYC'* or '*FAT4* drives cancer progression by stimulating cell proliferation and enhancing cellular invasion and migration' are considered acceptable is because *we might derive observable consequences that can be corroborated by selected experiments*. In other words, hypotheses must be consistent with observation, i.e. *empirically adequate*.

3. CONCLUSION

In this chapter, I elucidated the structure of discovery of data-driven biology. I discussed the cases of cancer genomics and genome-wide association studies showing how their structure fits very well the tripartite framework elaborated in Chapter 1. I have highlighted the role of background assumptions in hypothesis generation, and the eliminative inference structure of hypothesis development. Finally I have listed the epistemic values guiding each phase of discovery. Throughout the chapter, I have also sparsely discussed the important role that biological databases play both in theorizing and in scientific practice, especially in identifying biological phenomena and entities of interest in data sets. In the next chapter I will discuss in depth the heuristic contribution of databases, pinpointing exactly the epistemic role of biological databases in the discovery strategies of data-driven biology.

CHAPTER 3

THE USE OF BIOLOGICAL DATABASES IN SCIENTIFIC DISCOVERY

CHAPTER ABSTRACT

In this chapter, I analyze the epistemic role that biological databases might possibly play in contemporary biological research. In particular, I claim that there are two main ways of using databases. First, databases are used as 'evidence-enhancer', i.e. the play a vital role both in elaborating claims about phenomena and in strengthening the case for data as being evidence for a phenomenon. Next, databases are explored with theoretical aims in mind. Through data mining, databases are explored to identify robust patterns of data in order to compile what Lindley Darden calls 'the store of a field'.

INTRODUCTION

In Chapter 2, and in particular in the section on hypotheses development, I sketched an analysis of how, by means of comparison with data sets of biological databases, biologists might infer the existence of certain features/biological entities in their own data set. In this chapter, I develop this issue. How exactly do data sets of biological databases provide evidence for features of other data sets? What is the epistemological motivation supporting the reliability of this operation? Does this use of biological exhaust the epistemic role of databases in contemporary biology?

In this chapter I claim that there are *two main ways of using biological databases* in contemporary molecular biology that should be distinguished because they serve different purposes. First, biological databases are *evidence-enhancers* in the sense that they play a prominent role in strengthening the case for data as being evidence for a phenomenon or for claims about a phenomenon. Second, biological databases are *explored* with theoretical aims in mind. In the section on hypothesis development of data-driven biology, I showed that certain databases are *mined* in order to provide eliminative principles for hypothesis development, thereby transforming standard databases into *maps*. Here I argue that such operation of data mining, which I call *exploration of databases*, is an instance of *exploratory experimentation* (Waters 2007) and it is aimed to develop what Darden (2006) calls 'store of a field'.

This chapter should be conceived as an analysis complementing what I have already shown about data-driven biology in Chapter 2: Since databases are pervasively used in the discovery strategies of data-driven biology (but also in part in traditional

molecular biology), elucidating the epistemic moves that they allow is part of the analysis of the discovery strategies of contemporary biology.

1. DATABASES AS EVIDENCE-ENHANCERS

In this section I analyze how databases aid biologists (a) to infer the existence of phenomena in their data sets and (b) to elaborate claims about phenomena. My argument relies on the fact that data stored in databases are not data, but rather *phenomena type* (Teller 2010). By transforming data into phenomena type, databases enable the transformation of data into a basis of comparison for elaborating claims about biological phenomena and for identifying phenomena across different experimental contexts. Before delving into my main argument, I briefly specify what I mean by "data" and "phenomena", and the kind of philosophical literature I refer to (1.1.1.). In section 1.1.2 I discuss Leonelli's (2009) position and I will argue that her analysis, at least for the use of primary databases, is unsatisfactory.

1.1 Data, phenomena and claims about phenomena

In an influential paper (Bogen and Woodward 1988) and several restatements and defenses (Woodward 1989; 2000; 2011; Bogen and Woodward 1992; 2005) Bogen and Woodward put forth a three-level picture of scientific theory (data, phenomena, explanations) based on the distinction between data and phenomena.

Data have to be conceived as public records produced by measurements 'in the wild' as well as measurements within experiments. For instance, Hacking defines data as "uninterpreted inscriptions, graphs recording variation over time, photographs, tables, displays" (1992, p 48).

Data are evidence for *phenomena*. Phenomena are understood as features of the world that in principle can occur under different context or conditions – they are detected somehow as patterns or regularities[23]. Data then serve as evidence for both identifying phenomena and elaborating claims about phenomena.

Claims about phenomena take the form of *systematic explanations*. According to Woodward (1989) good explanations must meet, at least, two requirements. They have to exhibit the details of the *patterns of dependency*. Indeed, an explanation of the form

[23] The word 'phenomenon' has a long and rich history. Hacking provides a close examination of this history in (1983). However, here I restrict my use of the word 'phenomenon' to the meaning outlined in this section

'when the gene x is overexpressed, so is the gene y' is intuitively unsatisfactory. A systematic explanation needs to show in detail how the features of the explanandum (the phenomenon) depend upon factors appearing in the explanans. In molecular biology what is required is a detailed account of the causal mechanism producing the phenomenon (Darden 2006). A second feature of good explanations is their ability to systematize and unify; claims about phenomena should make reference "to factors, generalizations, or mechanisms which can figure in the explanation of a range of different phenomena" (Woodward 1989, p 400). Therefore scientists are not interested in data *as data*. Rather, they are interested in data to the extent that they exhibit some 'phenomenal' feature.

Data and phenomena differ in many respects. In particular, phenomena exhibit similar features in a wide range of situations, while data occur in the way they occur just in the particular experimental context in which they are generated. Gathering evidence for phenomena is challenging since we have to distinguish features of the phenomenon from features of the artifact (Woodward 1989, p 397). In other words, phenomena are 'evoked' only in a particular data set with some experimental moves, and data would reflect also the particular and specific conditions of the experimental context[24].

If data depends on the experimental context where they have been generated, the evidential scope of data is limited – *limited to the experimental context in which it has been generated*. Leonelli (2009) captures this idea by saying that data are *local* while claims about phenomena are *nonlocal*. She takes a fact to be local when the evidential scope of that fact strictly depends on the context where it has been generated, while a fact is nonlocal when its evidential scope goes beyond that context. Data are local in the sense that researchers have to be aware of the experimental context of data in order to establish which claims their data set supports. The scope of data does not go beyond their experimental contexts, although the features of the phenomena they show might be found in other situations.

However, data sets could be combined. By combining data sets, these become 'nonlocal'. Actually, there are two respects in which data can be taken as 'nonlocal'. Consider the following two statements:

1. The data set *x* and the data set *y* can be combined to enhance (or refine) a claim about a phenomenon *z*
2. A data set *x* can enhance the evidence for the presence of the phenomenon *z* into the data set *y*

[24] Famously, data are "idiosyncratic to particular experimental contexts, and typically cannot occur outside of those contexts" (Bogen and Woodward 1988, p 317)

As far as statement 1 is concerned, if we detect the features of a phenomenon *z* from a data set *x*, then we can formulate a claim about the phenomenon *z* on the basis of *x*. If a data set *y* is available, and from *y* we infer another feature of *z*, then we can elaborate another claim about *z*. We can combine the results of *x* and *y* in the sense that we can elaborate a claim about *z* that contains the features of *z* inferred from *x* and *y*. In the sense of statement 2, a data set is used as evidence for detecting a phenomenon in another data set.

Biological databases are used not only in the sense of statement 1, but also in the sense of statement 2. In the next sections, I will show how primary databases realize statement 2, and how metadatabases accomplish statement 1. In both cases, *data stored in biological databases are used beyond the initial experimental context where they have been generated*, i.e. data stored in biological databases are *nonlocal*

1.2 Leonelli's proposal

Sabina Leonelli offers an explanation of how biological databases expand the locality of data by describing how curators of databases standardize data sets (Leonelli 2009; 2010; 2013; Leonelli & Ankeny 2012). Leonelli proposes that curators of databases are able to *expand the evidential scope of data*. Curators have the aim of integrating data into a common framework. She calls this strategy "packaging data". This has the goal of transforming data in a way that can also be re-used as evidence for other phenomena.

Packaging of data in a biological database requires (at least) two main activities. First, database curators gather data from publications or other databases (Leonelli and Ankeny 2012, p. 31). Data are the targets of the curators of databases. However, curators of databases are interested in just a small fraction of data that can be found in scientific articles or other databases. Curators strip away what is irrelevant to the topic of the database. For instance, if one is curating a database on DNA methylation, then she will select just those data in articles that are relevant to methylation.

After selecting relevant data, curators annotate these through classificatory systems that aid the retrieval by users (Leonelli and Ankeny 2012, p. 31). Examples of labelling systems are the so-called *bio-ontologies* (Boem 2015), defined as formal systems of terms denoting biological entities or processes (Leonelli 2011, p. 333). Bio-ontologies connect phenomena for which data function as evidence with "the terms used to formulate claims about those phenomena" (Leonelli 2009, p 742). Moreover, in order to grasp the relevance of data produced by someone else, databases should also include information about the production of stored data because, as emphasized above, these two processes affect the kind of data obtained. Thus, a user might be interested in seeing

whether the data production and interpretation tools being employed are compatible with the ones used to generate the small facts stored in the databases. For these reasons, curators devise another type of label for data called 'metadata[25]' (Leonelli 2010, p. 336). This means that a small fact is categorized according to the way it has been obtained (e.g. ISS, inferred from sequence similarity[26]).

Through these classificatory systems and labels, curators *de-contextualise* and *re-contextualise* data at the same time (Leonelli 2011). De-contextualisation makes data "extremely adaptable, […] by stripping them of as many qualifications as possible, leaving them free to travel as objects in search of a new interpretation" (Leonelli 2011, p. 338). However, by means of metadata, data are *re-contextualised* at the same time and hence, a user can assess the provenance of data at any time. With the operation of de-contextualisation curators free data from their locality, opening up possibilities for the re-use of data in other contexts. However, with re-contextualisation curators provide enough information to shed light on the specificity of data's locality, thereby helping the user to re-use that data set in the proper context.

In my interpretation, the evidential scope of data is not exactly enhanced according to this explanation. In Leonelli's framework, data stored in databases are standardized according to the categories of a specific bio-ontology, but curators somehow provide information about the locality of data themselves. This information is useful for users in order to understand whether the locality of the experimental context of the users is compatible with the locality of the experimental context where data (stored in databases) have been generated. Therefore data stored in databases can be used in other contexts only if their original context is compatible with the new context where data need to be used. In other words, databases do not enhance the evidential scope of data, but they aid the recognition of compatible localities.

1.3 Data, phenomena types and biological databases

Leonelli's explanation pinpoints how curators' procedures aid the identification of compatible experimental systems. Here I want to show how databases enhance the evidential scope of data. The starting point of my argument stems from some considerations about the nature of data stored in biological databases. Let us start from a simple case of re-use of data stored in biological databases.

[25] Gene Ontology calls these labels 'evidence codes'.
[26] 'Inferred from sequence similarity' means that the annotation has its basis on a sequence-based analysis. For example one may infer the function of a gene by the function of one of its orthologs.

Imagine that one wants to do a *knock-in*. Knock-in refers to the technique of the insertion of a coding cDNA sequence into a genomic locus of interest[27]. In order to see whether the insertion has been successful, one sequences the DNA of the sample where the gene has been inserted, and checks whether the gene is actually in the specific locus. One way to verify the insertion is to use BLAST (*Basic Local Alignment Tool)*, i.e. an algorithm designed to compare primary sequences of protein as well as DNA sequences. BLAST is connected to several repositories of nucleotides sequences. Through this tool one may compare the sequence of interest obtained through sequencing technologies with a repository of sequences stored in multiple databases.

What exactly does one look for when comparing a sequence to a whole database of sequences by means of algorithms such as BLAST? In the simplest case of trying to see whether a gene of interest is actually in the selected genomic locus, one wants *to identify a biological phenomenon* in her data set. If phenomena are to be understood as features of the world that in principle can occur under different contexts or conditions, and they are detected as patterns or regularities (and hence they have some more or less stable features as Bogen and Woodward mean), then genes are phenomena. A gene *x* has sufficiently robust features (e.g. the sequence of nucleotides, though might slightly change, is quite uniform in a given species; if transcribed, can generate sufficiently robust gene products such as mRNA with a specific RNA sequence; etc) to count for a phenomenon. Therefore one compares a sequence to a library of sequences of a database by means of BLAST in order to establish whether the sequence she has found out is sufficiently similar to the established sequence of a specific phenomenon. Biologists, by using BLAST, are not interested in data (stored in databases) *as data*, but in a particular setting of data representing certain established features of a phenomenon.

My claim is that data stored in biological databases are not data proper, but rather phenomena. However, they are a particular type of phenomena.

1.3.1 Phenomena type and phenomena token

A distinction made in the literature on data and phenomena is particularly useful to explain what kinds of phenomena are stored in biological (primary) databases. This is the distinction between *phenomenon type* and *phenomenon token* (Teller 2010).

Let us consider the melting point of lead, i.e. 327.46 °C. Actually, there is no general melting point of lead 'out-there-in-the-world', in the sense that this threshold

[27] This technique is used in disease modelling to create, for instance, transgenic mice and to see the consequence of the action of a particular gene product in a controlled context. A similar procedure is called *transfection* that is, the deliberate insertion of nucleic acids into a cell. In general, these are methods for introducing foreign DNA into a eukaryotic cell, by means of viruses or other 'vectors'.

does not apply precisely in every circumstance. Depending on the particular environmental context and the particular causal factors at hand, the melting point of lead will change slightly, *but just a little*. This means that the melting point of lead is an *idealization*, in the sense that is a distortion or a simplification of a particular fact pursued in order to grasp some peculiar features of that fact that one will find in other similar facts. The idealized melting point of lead is called by Teller 'phenomenon type' while the occurrence of this type in the real world is called 'phenomenon token'.

Another simple example might be of some help. Consider the general fact that salt will dissolve when stirred into water. Consider then the specific event, *hic et nunc*, of salt dissolving when stirred into water. The former case is the phenomenon type of the phenomenon 'salt dissolving when stirred into water' because it represents the salt dissolving into water independently of the particular environmental context in which it happens. The latter is the phenomenon token of the type 'salt dissolving when stirred into water'. Therefore an individual event is a token of a phenomenon type if it falls under it, where the type isolates certain core stable and regular features. Moreover, phenomena tokens are idiosyncratic to the particular context in which they occur. In other words, the occurrence of the phenomenon token is tied up in the complex details of its context; while phenomena tokens are events in the real world, phenomena types are idealized types involving simplifications and shortcuts of that natural event. *A phenomenon token is an instantiation of a phenomenon type, and a phenomenon type is an idealization of a phenomenon token*.

But if a phenomenon type is an idealization of a phenomenon token, then what kind of idealization does Teller mean? He is not very specific. There are several types of idealizations (Cartwright 1983; McMullin 1985; Weisberg 2007; Weisberg 2013; Elliot-Graves and Weisberg 2014) and my claim is that two specific types of idealization are used in different types of databases. In primary databases, biologists 'store' phenomena type idealized in a *similar* way as Galilean idealizers do (though with some remarkable differences), while in metadatabases a variety of minimalist idealization is put in place. In the case of primary databases, data sets are used to enhance the evidence for the presence of a phenomenon in other data sets, while in the case of metadatabases, data sets are used to refine claims about phenomena inferred from other data sets.

1.3.2 Biological databases, phenomena types and Galilean idealization

My main claim in this part of the chapter is that biological databases store *phenomena types* rather than data. Phenomena types are idealized types of real world phenomena. In

primary databases[28], phenomena types are idealized, at first glance, in the Galilean sense. However, Galilean idealization can only in part make sense of the phenomena types stored in primary databases.

In Galilean idealization a scientist represents a complex phenomenon of the world in a simplified and/or distorted manner in order to make the system more tractable[29]. Having obtained an adequate model, one can de-idealize (by eliminating the effects of the distortions) and go back to the target system (Weisberg 2007). Despite the distorted representation of a target system, in the long run Galilean idealization aims to give a complete and non-distorted representation of the phenomenon under scrutiny. Once one has a model (idealized in the Galilean sense) of a target system, the next step is to eliminate gradually the effects of the distortion to see if the model has captured 'the essence' of the real world phenomenon. Galilean idealizers aim at what Weisberg calls *the representational ideal of COMPLETENESS.* While this characterization of idealization fits in some respects with primary databases, it also differs in other important features.

Consider the above example of gene knock-in. Once the gene is inserted into the cells of interest, one has to verify whether the insertion has been successful. One then sequences the DNA of the cells of interest, and aligns the results to the Human Reference Genome using BLAST. The phenomena type stored in the database of the Reference Genome are, for instance, sequences of nucleotides of genes or other biological objects of interest. The Reference Genome is a single consensus haplotype representing the sequence of each chromosome (Church et al, 2015) stored in a nucleic acids database. The first assembly of the human genome dates back to the work of the HGP consortium that 'collapsed' sequences from 6 individuals. The 19th rendition, named GRCh37 (GRC stands for 'Genome Reference Consortium) is a "mosaic haploid genome derived from about 13 people" (Editorial Nature Methods 2010, p 331). Since it is a mosaic haploid genome, *the Reference Genome is the sequence of no one*, i.e. it is the sequence of neither the donors from which it has been obtained, nor the sequence of any other person. Rather, it is an i*dealization*. It is a sequence representing a set of phenomena (e.g. genes, promoters, etc) according to specific criteria (e.g. sequences), which can be different with respect to the actual sequences existing in nuclei, as well as missing basically all the features of the environmental conditions that one may find in actual nuclei. This means that the Reference Genome is not incomplete – after all, it includes everything of the features of interest – but it is *false*.

The Reference Genome could be preliminary conceived as a Galilean idealization, meaning that it is a distorted version of the genome sequence of any single person one

[28] Primary databases are databases storing sequences of nucleotides and amino acids

[29] Galileo used this notion of idealization when formulating the law of free fall by imagining a frictionless surfaces (first distortion) as well as perfectly rounded spheres (second distortion) in order to unveil the mathematical essence of the law of 'free fall'.

may find in the real world. Since it is continuously updated, it could be said that builders of the Reference Genome aim to capture more and more details of the human genome, in particular through the increasing precision of the under-development sequencing technologies. In a very loose sense, we might say that builders of the Reference Genome aim at COMPLETENESS. However, *this is just the perspective of builders*. If we look at users, things get more complicated. In the literature on idealization, builders and users of idealization tend to coincide. Here, builders and users make use of idealization in distinct manners. The builder of the Reference Genome constructs a representation of the genome by following the 'regulative ideal' of getting closer to an accurate representation of real genomes. *The situation of the user is indeed different*. The user makes use of this idealized version of the human genome to identify specific phenomena *hic et nunc*. She is not driven by ideal of COMPLETENESS when using the idealized version of the human genome. In the case of knock-in, one compares the sequence of interest to the Reference Genome in order to see whether some sequence in the Reference Genome representing a certain phenomenon type (e.g. the sequence of a gene) are sufficiently similar to her own sequence. If this is the case, then it is possible to say that the sequence is a hint of a specific phenomenon (the gene one wanted to insert). Therefore, one checks whether the sequence *x* is a token of the phenomenon type gene *y*. In other words, one uses the Reference Genome to identify phenomena in the real world.

How do we establish how sufficiently similar sequences must be? Clearly this is a pragmatic choice. But the beauty of an instrument such as BLAST is that it provides a precise measure of similarity, which in biological jargon is called 'measure of *sequence identity'*.

For instance, imagine one wants to knock-in the gene *SETBP1.* One then sequences the DNA of interest and she obtains the following sequence:

AAGCGCGGGGCCGGCCGGGGCGCCCGCCGCTCGCCGCCGCCTCCGCGCGCCCGGGGGCCCCGGCGCCCCC
CGAGCTGGGGCGGCCAGCTCGCGGCTCGCCGTTTGACAGATGCTCATCGCCATGGAGTTGCCGCAGCAGCACCTTTGGGG
GCTCGGGCGAGCGACGGGAGCCGGGATCTGAGCGAGCGCCGGGGCCAGCGGAGCCGGAGCCGCCGGGACATGGTTGCA
GATCTGATCTCTTCTGAACACCTCATCGTGTCTCCATCCCTGGGAATCTGACCCTAGCAACTGGACCACTTTGTTCTTGGAAT
TTTGGGTGTCCTCTTTTCTCACCTTTCCCTTTTCCCTTTTCCCCTTCCCCCTCCTGAGAACTCCGGAAGACTGTAGAGATTGTC
ATGGAGTCCAGGGAAACCTTAAGCAGCTCCCGGCAAAGAGGGGGCGAGTCAGACTTCCTGCCGGTCTCCTCAGCCAAGCC
CCCAGCTGCTCCTGGCTGTGCAGGAGAACCTTTGCTCTCCACTCCAGGACCTGGGAAGGGGATCCCGGTGGGCGGAGAGC
GCATGGAGCCAGAGGAGGAGGATGAACTAGGCTCAGGGCGGGATGTGGATTCCAACTCCAACGCGGACAGTGAGAAATG
GGTGGCAGGAGATGGTTTGGAAGAGCAGGAATTTTCTATCAAGGAGGCAAACTTCACAGAGGGAAGTCTGAAGCTAAAGAT
TCAGACCACAAAGCGGGCTAAGAAACCCCCAAAGAATTTGGAGAACTATATATGTCCACCTGAGATCAAGATCACCATCAAG
CAGTCTGGGGACCAGAAGGTGTCCCGTGCTGGAAAAAATAGCAAAGCCACGAAGGAGGAAGAAAGAAGCCACTCCAAAAA
GAAGCTCCTCACAGCCAGTGACCTTGCAGCCAG
TGACCTCAAAGGATTTCAGCCACAGATTAAAGACTCCAGTAAGGAGGAAGTCTGGAAGAGAAGAGGAGGCCAAGGCATCCC
ATTCAAAAAGCAATTCCTGTCCCAGGAACGTGCCATGTGCTTCTCATGCCCCCGGAACCCATTCCCCGCAAAACCCGGTTCT
CTCACTCTTCCTTTTCACAGTGAACCTGCAGTCTGGGCACAAGAAGTATAACTTCGCATGGATTCTGCAAAGCCCACACCTGT

GGTCATTCCCTGTTCTTTCCATTCAACAATGGAGACTTGCCCAAGATTGTAAACTAGTGAGTGACAGCATTTGGGCTTATGAT
CTTCTCTGCCTGCTAGCTAGACATTCTCTCTGGGTCTAAAAGATAATCCAAAAAGATCCAGCTTCACAATGCTGCCCTGAAGA
GATAATGCATTAGGCGGCCCTGATGCAGCATTCACTGTCTTCCAGGGCAGGCTTGATCCCATGAGTTTGCTTGCTCAGACGA
TCACTTAGGAAACACATGCCTTTACTCTCTAGGCCTTTTTTCTAGCTTGCCTTGTATCAGCTATGCCATGGATCTTTGCTCTCT
TACCCCATGCTATACAGAGTATGGGCTCCAAGCCACAGCTGGCCTGTCAAGTGTGTGTCGCTGGTCCACCATGGGATACATT
TAGAAACTTTTATAGCAATTTGACATTTTTGTGATATCCAAGCATGTGATTGTTTTCCTACGGATTTGTCTTATAGTATTTTACC
AAAGTTTCCACACAAAAAGTATGGATTAAGGACAAAGTATCTGGTCCTTCATCAAAGATCGTTTGATAAGCTCTGTTCTAGTT
AACCAACACTGAGCTTTCCTAGTTTTAATAAAAGAGTAGGATTTGGAAAAAAAAAAAAAAAAAAA

Then, she runs BLAST, she compares that sequence with the Reference Genome and she obtains the result shown in Figure 3.1.

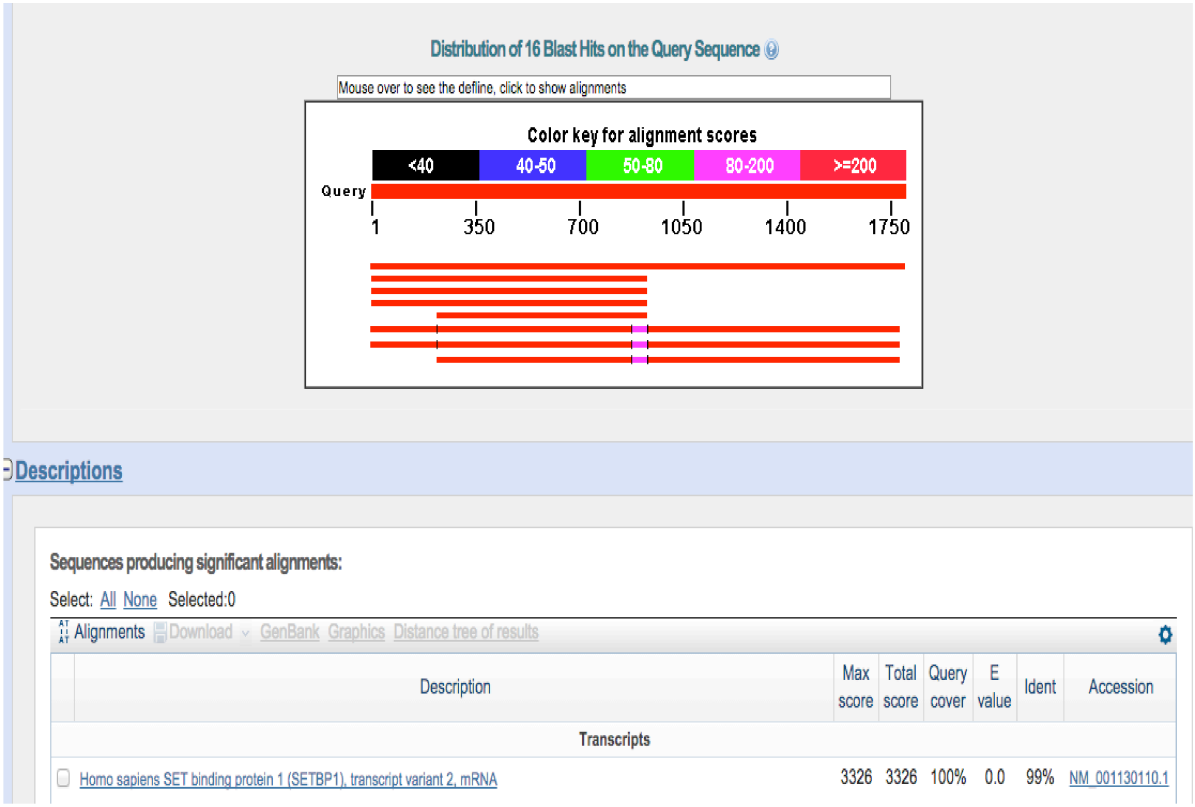

Figure 3.1. A screenshot of a BLAST's results page. At the bottom of the screenshot the reader might note the 99% of identity of the query sequence with the stored sequence of the gene *SETBP1*

Now, there is a 99% identity of the query sequence to the sequence of *SETBP1*. For a biologist obtaining a 99% match means that her sequence is sufficiently similar to the one found in Reference Genome to claim that her sequence is a phenomenon token of the phenomenon type *SETBP1*.

Therefore, the way users use the Galilean idealization of the Reference Genome is quite peculiar. Usually, Galilean idealization is used to make tractable a non-tractable problem. Once you have idealized and you get some sense of how the idealized model works, then you eliminate the idealization with the aim of going back to the target

system. The model then is a good model if it is similar enough (in certain respects) to the target system, in the sense that once you add details to your model, this fits the real world phenomenon in the respects you are interested in. For the user of primary databases the situation is reversed. It is the real world phenomenon (the phenomenon token) that should be similar to the 'model' (the phenomenon type). Let me be clear on that.

A model is a "good" model (a representation is a "good" representation) if the model (representation) is sufficiently similar (according to some criteria) to what is modeled (represented). The idea is that a phenomenon token is a good phenomenon token if it is sufficiently similar to a phenomenon type. Better, *one has identified a phenomenon if the supposed phenomenon is sufficiently similar to a phenomenon type*. In other words, identifying phenomena in the immense sea of data in genomics – at least in sequences alignment – depends on a similarity relation with respect to an idealized type. Therefore the evidential scope of data stored in biological databases goes beyond their locality and applies also to other experimental contexts (at least in the case of Galilean idealization) because data in databases *are phenomena type, and they are used as (one of) the condition sine qua non for establishing evidence for phenomena in other experimental contexts*. In other words, a primary database (in this case the Reference Genome) enhances the evidence for the identification of phenomena in other data sets because it is a repository of phenomena types *establishing the space of possible phenomena one may identify in wet labs data sets*.

For the user, the path then is not to start from the real world as a benchmark to get an idealized type and finally to go back to the real world, but to start from the idealized type as the benchmark to get a token; the direction of fit is different from traditional Galilean idealization.

1.3.3 Biological databases, phenomena type and minimalist idealization

Data stored in biological databases can be also phenomena type idealized in the *minimalist sense*, though with a caveat. Minimalist idealization is, in Weisberg's own words, "the practice of constructing and studying theoretical models that include only the core causal factors which give rise to a phenomenon" (Weisberg 2007). This type of idealization aims to identify the core causal factors producing a phenomenon[30]. A typical

[30] Minimalist idealization can take various forms and it can be linked to causal accounts of explanation. In a sense, minimalist idealization coincides with the notion of abstraction (Floridi 2008; Cartwright 1989; Ratti 2014), that is, "we strip away (…) all that is irrelevant to the concerns of the moment to focus on some single property or set of properties, 'as if they were separate'" (Cartwright 1989, p 187). Though Galilean idealization and abstraction might coincide sometimes, they clearly put in place two different operations. Galilean idealization is a deliberate distortion, while abstraction is an omission. Galilean idealization is false, while abstraction is simply incomplete. Moreover, minimalist idealization does not aim to COMPLETENESS. Strictly

molecular biology representation of a mechanism with arrows and boxes is a minimalist idealization because it contains only the (claimed) core causal factors producing the behavior of a particular biological system. The caveat is that databases can store phenomena type minimalistic idealized in the sense that they characterize phenomena in terms of features that could potentially make a contribution to causal stories. In this sense, databases aid to build and refine *claims about phenomena*.

If we assume a causal account of explanation, then minimalist models aid (and drive) scientific explanation. For instance, if we take Strevens' account of causal explanation (i.e. to give an explanation of a phenomenon is to give a causal history about why that phenomenon occurred, see Strevens 2004) then minimalist idealization becomes ubiquitous.

In certain cases, data stored in biological databases might be considered as phenomena type idealized in a minimalist sense. In particular, this is the case of data stored in *metadatabases* that is, databases organizing in a more convenient form data about a certain object of inquiry (e.g. cancer genes, methylation, etc). These databases store data about specific aspects of biological phenomena that can aid the identification of potentially core causal factors and that are relevant to causal explanations. Since metadatabases store data related to certain features of phenomena and not others, metadatabases idealize biological phenomena in a minimalist sense.

Let us consider the case of querying for a well-characterized cancer gene (*PTEN*) in the metadatabase Networks of Cancer Genes 4.0 (NCG4) that I have collaborated to put in place. NCG4 (An et al 2014) is a metadatabase collecting data about systemic properties of cancer genes. Figure 3.2 shows what one finds by looking for information about this gene in NCG4.

speaking abstraction is not really an idealization, but the pluralism put forth by Weisberg about idealization (that is widely accepted) tends to consider any modification on the representations of a target system for the purpose of better studying the target system itself as a case of idealization broadly conceived. I am aligned with this view.

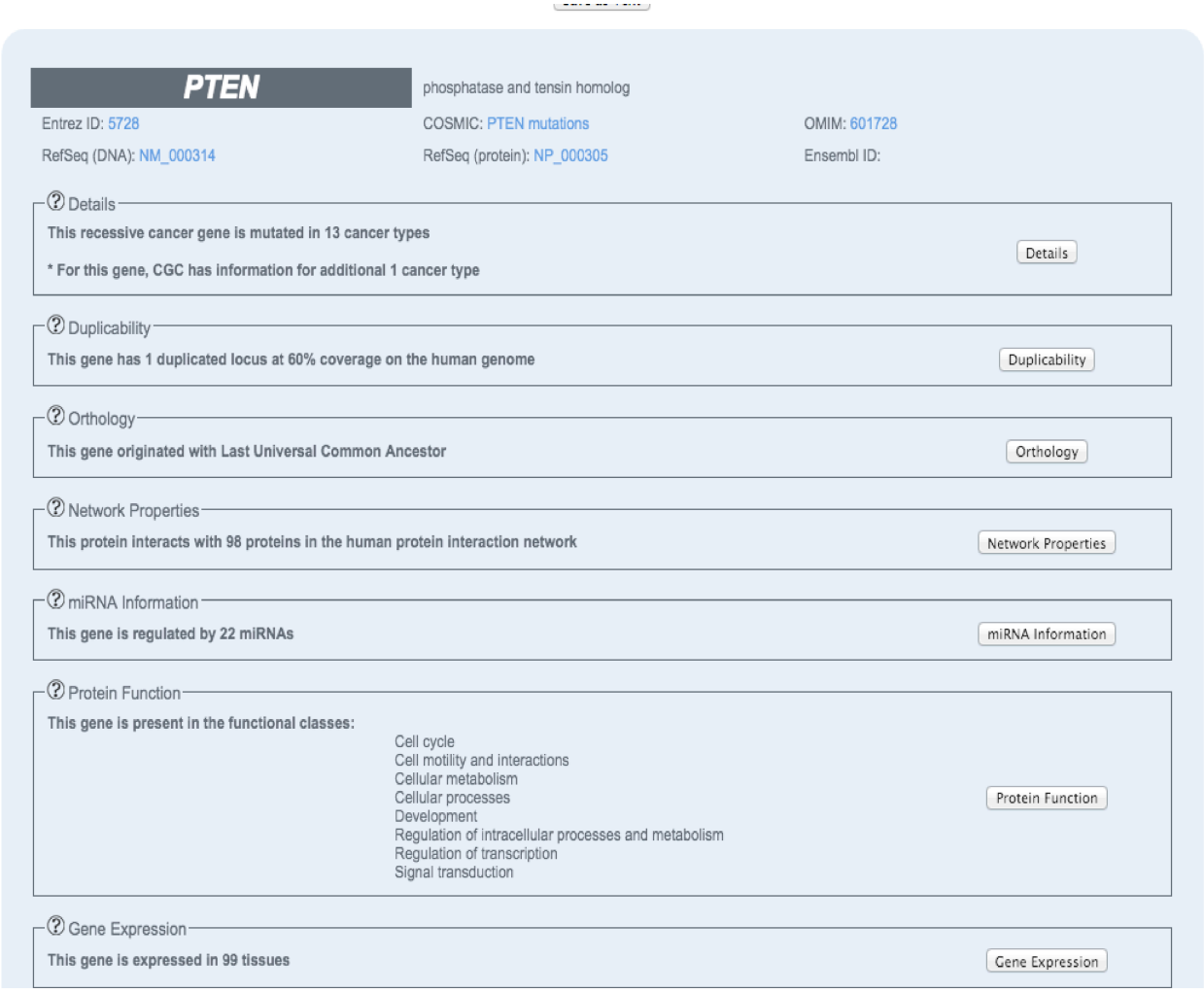

Figure 3.2. Screenshot of the results page for *PTEN* in NCG4

The broad aim of a database such as NCG4 is to provide key information about 'system-level' properties of cancer genes, i.e. genes playing a key role in cancer development. In other words, NCG4 provides information about some phenomena (genes). This information is interpreted data on the role of these phenomena in the development of cancer. *Sensu latu*, NCG4 provides information about the causal role of certain phenomena (genes) in other phenomena (varieties of tumors). For example in the section of NCG4's page 'details' (Figure 3.2) one might see that some articles provide evidence for the driver role of this gene in 13 tumors. Next, NCG4 provides also data about the origin of this gene in the Last Universal Common Ancestor. This is important since several cancer genes have a Last Universal Common Ancestor origin. Moreover, NCG4 collects data also about the functions of the proteins synthesized by *PTEN*. Data about origin, and function can be turned into information that aid the formulation of explanations in terms of a causal story of the gene *PTEN* being a driver gene in a particular cancer observed *hic et nunc*.

The representation of *PTEN* provided by NCG4 is not just an aggregation of data. It is a phenomenon type idealized in a minimalist sense. It is a phenomenon type because the way it is represented does not correspond precisely to any of the phenomena *PTEN* one might find in the real world. It is a minimalist idealization because NCG4 provides only the information useful to fill black boxes in a causal story where *PTEN* might be involved, i.e. the phenomenon is represented by abstracting some features of interest. How does exactly a tool such as NCG4 aid the construction of causal story and in general claims about phenomena? Imagine a typical cancer genomics screening. Imagine that certain genes turn out to be significantly mutated (according to the BMR). The next step is to hypothesize which genes are likely to play a role in the development of cancer. Here comes the minimalist idealization provided by metadatabases. One could check NCG4 to see the functions, the origins and the networks of interactions of the genes that, from a statistical point of view, are more likely to be driver. For instance, Lawrence and colleagues (2013) compare the phenomena tokens they have identified (genes recurrently mutated) to their types - *idealized in the minimalist sense* (the functional characterization) – and they choose not to consider certain genes as being driver. This is because the minimalist idealization – by characterizing genes in terms of their contribution to causal explanations - excludes these genes from a possible causal story since their functions have nothing to do with cancer development (or, at least, they have nothing to do with what is so far known about cancer development). Consequently, people can hypothesize whether certain phenomena tokens they have identified might play a causal role in the cancer they are dealing with. In other words, metadatabases – by minimally idealizing data as evidence for phenomena – enhance the locality of data, thereby allowing their re-use across several and different context.

1.4 Concluding remarks on the use of databases as evidence-enhancer

In the first part of this chapter, I depicted the evidential role of databases. I showed how databases store phenomena type and not properly data, and that these phenomena type might be used as evidence to claim that one has identified in her own data set a specific phenomenon as well as to aid the construction of claims about phenomena; primary databases are subjected mainly to Galilean idealization, while metadatabases are prominently repositories of phenomena type idealized in a minimalist sense. Now I am going to sketch an analysis of another use of databases: Data mining, namely the exploration of biological databases which purpose is not completely clear.

2. Exploring databases

There are many studies in big data biology that are not data-driven but are actually part of the heuristic strategies of data-driven biology itself. These studies use biological databases, but databases are not treated as evidence enhancers. Rather, they are increasingly used with a different purpose in mind. The studies I am talking about are a direct consequence of the emergence of big consortia like The Cancer Genome Atlas (TCGA), the ENCODE project or the 1000 Genomes Project. Some projects are designed explicitly to be maps (such as the ENCODE Project or the HGP), while others become maps only derivatively (e.g. TCGA). By joining forces from single scientific labs, big consortia are able to generate far more data than a single scientific lab. For example TCGA has sequenced, up to 2013, the genomes and the exomes of more than 3,000 cancer samples (Ciriello et al. 2013). The amount of data generated by the 1000 Genome Projects is suggested by its name. ENCODE has recently characterized the biochemical activities across the human genome's regions of several human cell-lines (The ENCODE Project Consortium 2012). Databases store these enormous data sets and, in the last few years, computer scientists have started to look for patterns in them. It is easy to find in journals like *Nature* or *Science* studies characterizing trends and patterns found in the data sets of TCGA, ENCODE or similar big science projects. Terms such as 'comparative analyses', 'system-level characterizations', and 'emerging landscapes' have become keywords. For instance, in the last two years the TCGA's immense data set has attracted much attention and it has become the object of inquiry of many mining studies. Consider the study by Ciriello et al. (2013). The wealth of genomic data generated by TCGA, Ciriello and colleagues say, is an *unprecedented* opportunity to *systematically* analyse similarities, differences, patterns, and signatures intra and inter tumour types. The idea is to compare and to organise data and to find regularities that would not be detected with few samples. By grouping 3,299 tumours from 12 cancer types Ciriello et al find out (a) a trend that divides tumours into two big classes, one characterized by somatic mutations and the other characterized by copy-number variations (CNVs) and (b) within each major class, there are *specific* oncogenic pathways altered (Figure 3.3). The aim of the article is to reduce the complexity of thousands of molecular alterations discovered in thousands of tumours to a few hundred types and patterns, and categorize tumours on this basis. There are many similar studies. Some focus specifically on copy-number variations (Lì et al. 2012; Zack et al. 2013; Kim et al. 2012) or on somatic mutations (Kandoth et al. 2013), while others on mutated genes (Tamborero et al. 2013) or on the analysis of trends in the functional annotations of the human genome (The ENCODE Project Consortium 2012). I call these studies *mining studies* and I claim that they constitute the second main use of biological databases: *The exploration of databases*.

Despite the growth of this kind of studies, neither their purpose nor what kind of practice they represent is entirely clear. In the next sections, I argue that these studies are used to form the kind of biological knowledge that forms eliminative principles. Moreover, the knowledge produced by mining databases can be used also to compile what Darden (2006) calls *the store of a field*.

2.1 The Structure of Mining Studies

The structure of these studies is straightforward. Roughly, a data set is explored with respect to data falling under one or more meta-data. A meta-data is a label 'attached' to a particular bit of data (or to a data set), and it is used to describe the datum itself. In other words, the metadata states what data is about. The idea of mining studies is that researchers look for patterns (i.e. robust regularities) in meta-data associations. For example, Kim et al (2012) mined the TCGA database looking for copy-number variations (CNVs). However, they do not look for regularities of CNVs with respect to their position along the human genome. Rather, they look for regularities with respect to another metadata, i.e. *tumour type* (defined by tissue of origin). Then, for each set of CNVs in each tumour type Kim and colleagues look for the genes located within the region amplified or deleted. These kinds of studies capture the disrupted pathways (and related disrupted functions in the cell) caused by CNVs alterations in thousands of tumour samples. By doing this, it is possible also to capture certain regularities and to say, for each tumour type, which are the biological processes that one might reasonably expect to find disrupted. In other words, it is possible to formulate 'generalizations' like 'in lung cancer, CNVs deregulates the pathways *x*, *y* and *z*'. Ciriello and colleagues (2013) provide similar generalizations when they discover two macro-categories of tumours (one with somatic mutations and one with CNVs) and that each tumour type (again defined by tissue of origin) falls, more or less, either under the category of tumours with somatic mutations or under the category of tumours with CNVs. To sum up the structure of mining studies, despite the complexity of computational tools used and the astonishing amount of data analysed, is simple:

(d) Scientists look for associations between different metadata labels in order to uncover macro-regularities
(e) Macro-regularities (i.e. patterns) are, strictly speaking, *predictions* in the sense that they provide an expectation of what it is likely to be found in similar contexts

2.2 Differences between data-driven studies and mining studies

Mining studies, despite being instances of big data biology, share few features with the data-driven biology screenings I described in the previous chapter. Data-driven biology and mining studies have in common the existence of background assumptions. However, *in mining studies background assumptions play a substantially weaker role than in data-driven biology*. In mining studies, the only background assumptions present are:

1. The theoretical basis of the computational tools used to identify associations
2. The fact that pattern discovery is metadata-laden, meaning that it is possible to find associations only within the categories of a pre-existing taxonomy system *x* (e.g. Gene Ontology).

In order to better understand what 'metadata-laden' means, imagine that the taxonomical system in which Ciriello *et al.* embedded their research classifies DNA mismatch repair and p-53 mediated apoptosis under the same label (Figure 3.3). Then, we would not be able to identify the patterns according to which DNA mismatch repair *is not* altered in ovarian cancer, while p-53 mediated apoptosis it is because we would classify DNA mismatch repair and p-53 mediated apoptosis as the same phenomenon. Therefore, it is the structure of metadata that provides the kind of patterns of data we can possibly detect.

Next, *in mining studies the ideal of discovering mechanisms plays no role*. What mining studies provide strictly speaking are *predictions*. While it is true that providing a mechanistic explanation also enables the formulation of predictions, the reverse is clearly false. Clearly I am assuming that, while mechanistic evidence is also a kind of statistical evidence, statistical evidence might not be mechanistic evidence. The fact that a SNP is shown to play a causal role in a mechanism that affects diabetes enables the formulation of the prediction that, whenever I find the SNP, there is a high probability of finding diabetes. However, having merely the association of the SNP with diabetes provides no mechanistic evidence that can explain the correlation. Douglas in (2009) says that the relation between explanation and prediction is a functional one[31] in the sense that predictions are valuable because they force us to test our explanations. In the case of data-driven biology, the reverse is true: Predictions (the hypotheses survived to eliminative induction) provide the cognitive path to mechanistic explanations. This can be true also for mining studies. The association of ovarian cancer and alteration (through CNVs) of p-53 mediated apoptosis is a prediction that also suggests an experimental path

[31] Actually, "explanations provide the cognitive path to predictions, which then serve to test and refine the explanation" (Douglas 2009, 454).

to uncover mechanisms. However, what I shall argue in the next section is that the kind of predictions established by mining studies does not suggest *directly* a path to uncover mechanisms. Rather, they provide something subtler.

2.3 Mining studies elaborate generalizations

What is the role of predictions provided by mining studies? What kind of role do they play in contemporary biological research? There is not much literature in philosophy of science about predictions as there is, for example, of scientific explanations. Actually, most of the literature on prediction is 'explanation-laden', and it is focused on the differences between explanation and prediction (Douglas 2009; Hanson 1959; Scriven 1959).

However, predictions may be seen also as *generalizations* (Shmueli 2010). The reason is that the function of scientific generalizations "is to provide reliable expectations of the occurrence of events and patterns of properties" (Mitchell 1996, S477). Hence, generalizations are somehow predictions.

Generalizations uncovered by mining studies play, in my opinion, a subtler role than traditional predictions or expectations. My thesis is that generalizations derived from mining studies *might provide some of the eliminative principles used to narrow the universe of hypotheses generated in the early phases of data-driven biology*. Let us see how.

Consider GWASs. Above, I have said that one late eliminative step is to provide a preliminary functional characterization of SNPs by looking at ENCODE data (Germain et al 2014). If a SNP either does not overlap to a functional region or it overlaps with a region which function is not related to the phenotype of interest, then the SNP is eliminated from the universe of hypotheses. Thus, ENCODE annotations - by making generalization on the functions of human genome's regions - have provided a way to develop eliminative principles to narrow the universe of hypotheses of GWASs. Let me provide an example. RegulomeDB[32] (Boyle et al. 2012; Schaub et al. 2012) is a public database guiding the interpretation of regulatory variants in the human genome. The database integrates many sources of functional annotations, including ENCODE data. Each variant is classified according to a score. The score provides to the user a quantification of the confidence that a variant is likely to influence transcription factor bindings, expression quantitative

[32] http://www.regulomedb.org/

trait loci, etc. For instance, variants that falls within category 1 are likely to affect protein binding and expression of a gene target while variants of category 2 are likely to affect only protein binding. Imagine that a researcher makes a GWAS for prostate cancer. She narrows the universe of hypotheses through the statistical eliminative principles illustrated above down to 2 SNPs. These are rs902774 and rs966321. ENCODE annotations found in RegulomeDB provide useful information for eliminating one of these two variants. The researcher through RegulomeDB will discard rs966321 because this variant falls in a region not associated to any biochemical activity. However, RegulomeDB provides important information for rs902774 that leads to formulate a specific hypothesis to be tested in the last phase of the discovery structure of data-driven biology. Not only this SNP falls in a functional region, but there is also evidence (score: 2a, i.e. the highest score for category 2 in RegulomeDB) that rs902774 falls in the binding site of the transcription factor CTCF and that it negatively influences the binding site of the transcription factor. Since the SNP is functional by influencing negatively a transcription factor (and this can have influences in the expression of genes regulated by CTCF), the researcher will prioritize rs902774 in the phase of hypothesis justification. In other words, through ENCODE annotations one is able to prioritize a SNP in particular, and to eliminate another from its universe of hypotheses. Hence, *ENCODE generalizations of the form 'this particular region of the genome is biochemically active in such and such a way' are able to provide useful principles to be used in order to interpret SNPs*. By analysing large cohorts of data, mining studies provide empirical generalization that can be used to prioritize certain hypotheses instead of others.

Similar considerations may be drawn for cancer genomics. As Raphael et al (2014) emphasize, the challenge in cancer genomics is to identify driver mutations and to understand their effects on pathways and cellular processes. The idea is that if certain genes are mutated, then (in light of functions and the pathways genes participate in) they might be good proxies for driver mutations. As emphasized in the literature (Raphael et al. 2014; Vandin et al. 2011) there are tools that, by grouping genes in terms of functions and pathways, may be of some help in restricting the universe of driver genes. However, how do we decide whether a pathway is relevant to cancer? A mining study like Ciriello et al. (2013) provides useful hints. Through the generalizations provided by Ciriello and colleagues, data-driven biologists obtain a list of cancer-specific pathways to be checked during the narrowing of the universe of hypotheses. Let us see this through a specific example. Consider Figure 3.3.

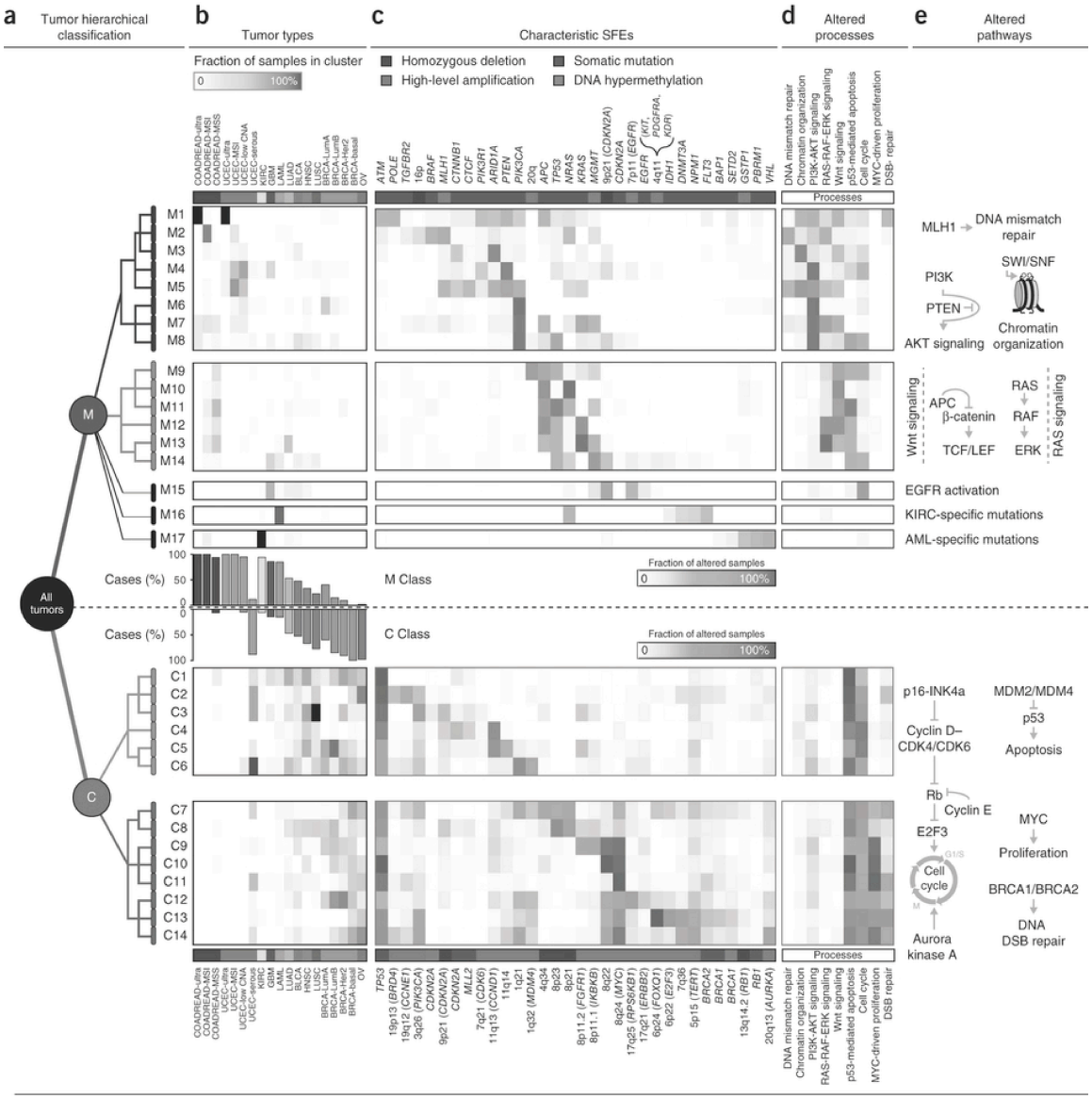

Figure 3.5. A visual summary of the patterns identified in the mining study (Ciriello et al. 2013). (a) Tumours are divided in two main classes: mutational tumours (M) and copy-number tumours (C). For M, Ciriello and colleagues have identified 17 subclasses, while within C they have found 14 distinct 'oncogenic signatures' . (b) Each tumour type (defined by tissue of origin) falls, mostly, in one of the subclasses . (c) Each tumour type has specific functional alterations related to specific genes . (d,e) Each tumour type has specific cellular processes and pathways disrupted. Shades of grey represents the fraction of samples in cluster. The figure is taken from (Ciriello et al. 2013) and it has been modified.

This figure is taken from Ciriello and colleagues' article. It is a visual summary of the patterns they have identified. Let me show how to use this 'map' by means of an example. Imagine that a physician has a patient affected by colon cancer (ultramutators variety). The physician then decides to genotype the tumour of the patient in order to grasp the genomic features of the tumour. A guide on what to look for in the genome is provided by Figure 3.3. First, if the tumour is colon cancer ultramutators (column b, COADREAD-ultra) variety, then it is a M1 tumour (column a). Next, somatic driver mutations should be located in selected genes like *ATM*, *APC* or *PTEN* and other genes that are located in the same pathways (column c). Hence the physician will look only

certain genes and not others (column c). Equally, in order to understand which are the altered pathways and altered cellular processes, the study of Ciriello et al provides an interesting source of criteria to prioritize only certain pathways and processes (in this case, chromatin organization, PI3K-AKT signalling, etc, as shown in columns d and e).

To sum up, the associations found by Ciriello et al might be considered as a pre-determined genes-and-pathways set that any researcher should compare with her list of mutations, genes and copy number variations. For instance, if certain genes do not overlap with the gene set provided by the generalization provided by Ciriello and colleagues, then they should be eliminated from the universe of hypotheses. In this sense, generalizations drawn through mining studies provide new eliminative principles or complement existing ones.

2.4 Mining studies, store of a field and exploratory experiments

In the previous section I argued that mining studies elaborate generalizations aimed at creating or complementing eliminative principles for data-driven biologists. Following this line of reasoning, we might say that mining studies are driven by a desire to find hints on how to look at an enormous amount of data when there are no specific expectations guiding observation. To put mining studies in a broader category of experimentation, we might say that mining studies are *exploratory experiments*.

For instance, exploratory experiments are "driven by the elementary desire to obtain empirical regularities and to find out proper concepts and classifications by means of which those regularities can be formulated" (Steinle 1997, S70). This is exactly the goal of mining studies, which aim to obtain patterns of data, extract generalizations and elaborate new classificatory frameworks. Exploratory experiments, Steinle goes on, emerge in periods of scientific development when a well-formed theory about certain phenomena is missing. Needless to say, the so-called 'big data' biology is still at its infancy and only recently scientists have started to uncover preliminary generalizations. Steinle also adds that exploratory experiments are not theory-free but rather they are somehow constrained by guidelines. Similarly, mining studies are constrained by metadata labels and the computational tools employed to discover associations of various sorts. O'Malley (2007) argues that exploratory experiments deal with complex interacting systems. This is the case of mining studies and explorations of genomes. As it is now widely shared, genomes are highly complex entities. Moreover, O'Malley adds that exploratory experiments constitute a broad inquiry based on multiple experiments and their relationships. As the examples above have shown, this is clearly the case of mining

studies. For example, mining the data-set of ENCODE means mining the relationships between data coming from 1,649 experiments.

If mining studies are exploratory experiments, and if they provide useful generalizations guiding scientific inferences, then mining studies seem to provide the terrain to form a sort of background knowledge for an entire discipline. They are like textbooks of a field (e.g. Alberts et al 2014), being a primary source for that field. In a sense, exploring databases through mining studies might be conceived as the practice of forming what has been called in the literature 'a store of a field' (Darden 2006). Darden, when talking about mechanisms as composed of both entities and activities, claims that "[f]or a given field at a given time, there is typically a store of established or accepted components out of which mechanisms can be constructed" (2006, p 51). In molecular biology, examples of components are genes or proteins. In the store of a field there are also accepted modules, which are organized complex of the entities of a field. Examples may include nucleosome, ribosomes, etc. The notion of a store of a field is not well developed but there are some features that can be straightforwardly identifiable. First, the existence of a store in a field puts important constraints on discovery strategies. If a molecular biologist has to discover a certain mechanism, the starting point of her research will be to consider certain entities and modules that she will combine to elaborate a mechanism description. But there are a limited number of types of entities and modules, which are exactly what form the store of molecular biology. Therefore, the way she will develop a mechanistic description depends on the nature of the store of her field – that is, the availability of specific entities, activities and modules. Another important feature is that these stores are usually given in a propositional form in textbooks like the famous one written by Bruce Alberts and colleagues (2014). In such textbooks practitioners describe, for instance, the biology of the cell and they describe how different entities interact to produce certain established biological phenomena, or how entities are combined to form specific modules. But there are also richer stores provided by leading reviews, which assume the kind of knowledge established in textbooks. In molecular oncology two examples of leading reviews are the famous series of reviews by Hanahan and Weinberg (2000; 2011) and the Vogelstein and colleagues' *Cancer Genome Landscapes* (2013). Hanahan and Weinberg's review provides constraints on how to build a mechanistic description of established processes related to cancer, like the evasion of apoptosis. Clearly, the description depicted by Hanahan and Weinberg are not taken to be exhaustive, in the sense that, while moving across the space of possibilities that they provide, one can add additional details thereby enriching the mechanistic description of the evasion of apoptosis. The review by Vogelstein and colleagues (2013) is another example of leading review compiling a store of a field. They

assemble the latest results in molecular oncology, a description of the main entities involved in cancer development, how they combine together to form modules, and so on.

But how exactly does one compile such a store? In the case of leading reviews and textbook one must scrutinize the literature, select accepted results, and organized them in a coherent framework.

My claim is that mining studies, and hence the whole practice of the exploration of databases, aims at building a store. However, this is a different kind of store. I would like to conceive stores with respect to the kind of constraints that they provide. In the case of textbook (as well as leading reviews), we may talk about *low-level constraints*, in the sense that textbooks provide just the *condition sine qua non* to elaborate mechanistic description in, for instance, the molecular biological field. Textbooks will tell you which are the minimal units to consider when exploring the biological realm. Then there are *medium level* constraints. These are modules: (again) textbooks and leading reviews provide scientists with a guide as to how combine entities in a certain way and not in another. Results provided by mining studies somehow assume low-level constraints (defined in the categories of, e.g., Gene Ontology) as well as medium-level constraints, and they provide an additional layer of constraints, which I call *high-level constraints*. By detecting patterns in big data sets, mining studies come out with a sort of regimented procedure for 'theory choice' (using this expression in a very broad sense) which guide discovery literally step-by-step as shown in the case of the physician analysing colorectal cancer. These high-level constraints limit the degree of freedom for the process of data analysis, thereby regimenting the whole process of discovery. In a sense, once you have certain initial features of the phenomenon you are analysing, the process of discovery descend almost consequentially with the use of high-level constraints. Given Figure 3.3, if one is analysing colorectal ultramutator variety, then the specific entities and activities - as well as modules and pathways - can be immediately identified, and not guessed among a space of possibilities by means of experiments. With high-level constraints, the space of possibilities is particularly tied.

Therefore, mining studies provide this sort of high-level constraints. While textbook or reviews provide stores by scrutinizing hundreds of articles, the practice of building a store of a field by means of the exploration of databases is achieved by means *exploratory experimentation*.

CONCLUSION

In this chapter I tried to make sense of the epistemic role of biological databases in contemporary research. I scrutinized their use, and how we can understand their use

according to the notion of idealization, which is central in scientific practice and theorizing. My aim was to explain from an epistemic point of view what the use of databases might allow and to fill a conceptual gap in the literature in the philosophy of biology.

Accordingly, I classified the uses of databases in two main varieties. First, databases are *evidence-enhancers*, namely that they are used to reinforce evidence for identifying certain phenomena in data sets and to aid the formulation of claims about phenomena. In another sense, databases are subjected to *exploratory experiments*, and they are mined in order to identify robust regularities leading to generalizations, which are then used as high-level constraints for building what Darden (2006) calls *the store of a field*.

CHAPTER 4

STRATEGIES OF DISCOVERY OF TRADITIONAL MOLECULAR BIOLOGY

CHAPTER ABSTRACT

In this chapter I identify the structure of discovery of what I call 'traditional' molecular biology. First, I identify clearly what I mean by 'traditional' molecular biology. Next, I analyze a strand of the existing literature on discovery in molecular biology, namely the so-called mechanistic philosophy. Next, I show how molecular oncology – a sub-field of molecular biology – follows the same strategies identified by the 'mechanistic philosophers' in molecular biology. Finally, I show how my tripartite framework fits molecular oncology and how the received view of discovery strategies in molecular oncology can be understood through my framework.

1 INTRODUCTION

In this chapter, I identify the discovery strategies of traditional studies in molecular biology.

This work is motivated by a controversy between scientists supporting data-driven biology as a novel set of scientific methodologies to uncover biological phenomena, and 'traditional' molecular biologists claiming that their 'traditional' strategies constitute (and will continue to be) the only right scientific method for biology. This is exactly a controversy over methodologies and discovery strategies. Therefore, in this work I am comparing data-driven and 'traditional' discovery strategies in order to see whether the controversy lies exactly at this level. After having discussed discovery strategies of data-driven molecular biology, I now turn to 'traditional' molecular biology. In particular, I shall be focused on molecular oncology and in general on the molecular studies on cancer stemmed from the development of the so-called 'oncogene paradigm' (Morange 1998). There are two reasons for doing this. First, in Chapter 2 my analysis concentrated on data-driven studies on cancer by scrutinizing case studies from cancer genomics and GWASs. Therefore, it would be appropriate to compare how different discovery traditions tackle the *same* type of phenomena. Next Robert Weinberg, who is arguably a pioneer in molecular oncology, has taken a prominent part in the controversy motivating this work. Therefore, by analyzing one of his latest works as a case study (Guo et al 2012), I discuss his conception of 'the right methodology for molecular biology'.

Before starting, it is better to grasp exactly what I mean by 'molecular biology', a concept that is far from clear.

1.1 What is molecular biology?

Defining molecular biology is a hard task. We might start from a very broad and common sense definition. One may say that molecular biology is the study of the biological realm at the molecular level. This is clearly too broad, and it seems to encompass much more than what people usually think. As Morange says, molecular biology cannot be just the description of biology in terms of molecules, because if this were the case, then even Pasteur would count as a molecular biologist (1998, p 1).

Some have proposed to see molecular biology as being unified by the notion of *mechanism*. To put it very simply, mechanisms "are entities and activities organized such that they are productive of regular changes from start or set-up to termination conditions" (Machamer et al 2000, p 3). According to Darden and Tabery (2009) the notion of 'mechanism' is the cornerstone to generate a clear and precise picture of molecular biology, at least from a philosophical (that is, conceptual) and historical point of view. However, this looks both imprecise and vague (in the technical sense of being susceptible to borderline cases). Strangely enough, Darden herself (Craver and Darden 2013) depicts the 'mechanistic tradition' as encompassing far more than traditional molecular biology. Actually, Darden and Craver say explicitly that the search for mechanisms is one of the biggest achievement of science, and that scientific activity as a whole (whatever this means) should be organized to advance mechanistic knowledge, and not only at the biological level. They also stress that this 'big project' dates back *at least* to the Scientific Revolution. The notion of mechanism, then, is not a peculiarity of molecular biology. Actually mechanisms are pervasive in basically *all* biological fields. From neuroscience to ecology, the project of discovering mechanisms is ubiquitous. Therefore we cannot say that the idea of discovering mechanisms unify conceptually and historically the field of molecular biology. Rather, one may say (though I do not agree) that mechanisms unify biology. But the notion of mechanism goes *well beyond biology*. For those who are unfamiliar with the history of science, even in the last few years there is much talk about mechanisms in physics, as for example the so-called 'Higgs boson mechanism'.

Another strategy is to argue that molecular biology is a *hybrid*. As many have argued, molecular biology seems to be the encounter of two different disciplines, i.e. genetics and biochemistry. However, it is not merely the sum of these disciplines. To quote again Morange, molecular biology is "a new way of looking at organisms as

reservoirs and transmitters of information" (1998, p 2). It seems that most molecular biological studies are focused on the notion of biological information (Floridi 2010), in particular on the code written in the DNA that then serves as a blueprint for making proteins. Information here is to be understood in a metaphoric sense. In particular, it is a metaphor helping researchers to give a rationale for conceptualizing a certain class of problems. Therefore we might say that molecular biology is a set of techniques borrowed from biochemistry and genetics, unified by a metaphor-heuristic, applied to the discovery of mechanisms at the molecular level.

However, this definition is not *historically* accurate. While it is true that early molecular biologists make use of techniques borrowed from genetics and biochemistry, it is also true that molecular biologists soon developed their own peculiar techniques. Therefore we should add a historical axis to our characterization of molecular biology. Fortunately, there is a consensus on the brief history of molecular biology. According to Morange (1998), the conceptual tools of molecular biology were forged between 1940 and 1965, while the 'consequent operational control' was put forth between 1972 and 1980 with the era of so-called 'gene technology' (Rheinberger 2007).

Therefore we may say that molecular biology is a discipline developed throughout three decades (1940s, 1950s and 1960s) from the combination of a set of techniques borrowed from biochemistry and genetics and unified by the informational metaphor-heuristic. In the 1970s and 1980s molecular biology matured with the development of gene technology. The whole set of techniques is applied to the discovery of mechanisms at the molecular level.

2 DISCOVERING MECHANISMS IN MOLECULAR BIOLOGY

After the characterization of molecular biology, the question is: Which are the discovery strategies of this discipline?

Gross (2013) concentrates his analysis of discovery strategies on the idea of (epistemic) complexity, a concept borrowed from Rheinberger (1997b). Epistemic complexity refers to the fact that biologists, when investigating a phenomenon, are usually in a situation of incomplete information and limited experimental access. This means that the complexity is not necessarily referred to the phenomenon itself, but rather to the scientific task. In a situation of limited access there can be a great deal of possible ways in which the phenomenon could be organized that are consistent with current knowledge. According to Gross (2013), asking questions about discovery strategies in molecular biology means asking questions about how molecular biologists reduce and attack epistemic complexity and more specifically how the search for mechanisms constitutes the way molecular biologists attack epistemic complexity. The

landmark work on mechanisms to be considered here is *Discovering Complexity* by Bechtel and Richardson (1993; 2010). Strategies of discovering in science are said to be mechanistic when scientific discovery is analogous to the attempt of explaining the working of an engineered machine. In a typical mechanistic explanation, relevant parts, relevant activities and their organizations are depicted and arranged into a chain of reasoning to show how parts and activities produce the functioning of the machine.

'Mechanism' should be understood as a metaphor grounding a set of heuristics. By '*heuristics'* I mean strategies or 'rules of thumb' (Langley et al 1987; Wimsatt 2007; Gross 2013) aimed at reducing the epistemic complexity of a scientific problem. Since an epistemically complex problem can have many possible solutions, heuristics strategies are used to reduce the list of solutions to that problem.

In the next sub sections of section 2 I will first characterize more in detail the notion of mechanism (Machamer et al 2000) and what does it mean to explain a phenomenon by discovering mechanisms (Bechtel and Abrahamsen 2005). Then I will introduce several strategies identified for discovering mechanisms[33] in particular Bechtel and Richardson's characterization (2010) and Darden's work (Darden 2006; Darden and Craver 2002; Craver and Darden 2013). Next, I will show these heuristics in action in the well-discussed historical case of the discovery of protein synthesis.

2.1 Mechanisms and mechanistic explanations

The concept of mechanism plays a key role in scientific activity. In molecular biology, the concept is used to describe (and explain) phenomena such as DNA replication, protein synthesis, cell differentiation, etc. However it is not entirely clear what mechanisms are or – better – to my knowledge there is no consensus among biologists about what exactly mechanisms are. It seems that thinking about mechanisms should include some discourse on relevant biological parts, their activities, and their organization. I will not fall into the metaphysical temptation of discussing whether mechanisms are real and what is the best way to characterize them; rather, I employ 'mechanism' as a metaphor-heuristic used to attack epistemic complexity in molecular biology. Therefore I will employ a particular notion of mechanisms that is instrumental to the strategies of discovery elucidated for molecular biology: Mechanisms "are entities and activities organized such that they are productive of regular changes from start or set-up to termination conditions" (Machamer et al 2000, p 3). Entities include macromolecules such as nucleic acids, proteins, and RNA molecules while activities include electrochemical activities as biochemical signaling, cell

[33] Actually, most of the debate on mechanisms is focused more on the *metaphysics* of mechanisms that is, what mechanisms are out-there-in-the-world. It would be interesting to show how the heuristic-metaphor has become a metaphysical program, but this task is well beyond the scope of the present work.

fusion, etc. Molecular biologists then attack epistemic complexity of biological phenomena by decomposing phenomena of interest in entities and activities in the attempt of finding out how these entities and activities produce phenomena themselves. One might say that talking about 'discovery of mechanisms' implies some sort of metaphysical commitment to the reality of mechanisms. However, here I refer to 'discovery of mechanisms' by using the word 'mechanism' just instrumentally (and hence without any metaphysical commitment), where 'mechanism' is merely something about the phenomenon of interest that could in principle *explain* the existence of the phenomenon itself.

Here comes *explanation*. Since the interest of science is to *explain* the natural world, and mechanism are taken to explain biological phenomena, does the discovery of mechanisms provide some sort of explanation of biological phenomena? There is a tendency among philosophers – originated in Salmon (1984) - to elaborate a specific notion of scientific explanation tailored around the concept of mechanism (see for instance Machamer et al 2000; Bechtel and Abrahamsen 2005). In particular, Bechtel and Abrahamsen point out that "the terms biologists most frequently invoke in explanatory contexts is *mechanism*. That is, biologists explain *why* by explaining *how*" (2005, p 422). Therefore mechanistic explanations explain in the sense that they show *how* a particular phenomenon is produced. In other words, a mechanistic explanation creates a causal story. Models then are explanatory when they describe mechanisms, i.e. "to give a description of a mechanism for a phenomenon is to explain that phenomenon, i.e. to explain how it was produced" (Machamer et al 2000, p 3).

But when exactly is a mechanistic explanation a good explanation? This should be linked to the notion of 'productive continuity' (Darden 2006; Darden and Craver 2002). In a nutshell, a good mechanistic explanation must show how *each stage* of the mechanistic chain produces the next one, i.e. improving the quality of an explanation means filling gaps in a chain connecting set-up and terminations side of the causal chain.

2.2 Strategies for discovering mechanisms (i): decomposition and localization

As many commentators argue, the concept of mechanism provides a heuristic to tackle scientific problems. This concept lays out the task involved in the sense that

"the scientist must identify the working parts of the mechanisms, determine what operations they perform, and figure out how they are organized so as to generate the phenomenon" (Bechtel and Abrahamsen 2005, p 432)

According to Bechtel and Richardson (2010) there are two basic mechanistic strategies for discovering mechanisms: *Decomposition* and *localization*. Again, these are mechanistic in the sense that they are understood in analogy with the functioning of an engineered machine. But decomposition and localization are instances of two more general strategies to isolate components of a system: Analytic and synthetic strategies. *Analytic strategies* identify relevant components of the system that are likely to play a non-negligible role in the production of a certain phenomenon by, for instance, intervening physically on those parts to determine what each part does. Analytic techniques include inhibitory strategies (i.e. determine the contribution of a system's part by inhibiting its activity) and excitatory studies (i.e. extra stimulation on an part's activity to determine an identifiable surplus). *Synthetic strategies* start with the hypothesis that the whole functioning of a system might be performed by a set of component operations. After that, one tries to assign operations to specific parts of the system. Synthetic strategies are used pervasively in the field of Artificial Intelligence (AI).

Analytic and synthetic strategies are complementary in the sense that some analytic strategies (such as inhibitory techniques) can provide evidence for certain synthetic models, while synthetic models might provide a framework to interpret analytical studies. According to Bechtel and Richardson (2010) one prominent way of interplaying analytical and synthetic strategies – especially in the discovery of mechanisms – is the combination of *decomposition* and *localization*.

Decomposition is a kind of heuristic allowing the decomposition of a complex explanatory task into a set of more manageable tasks. In particular, it assumes that the whole activity of a system is the result (the sum) of a set of subordinated functions. In this sense, decomposition is genuinely synthetic. Localization then is the assignment of different activities to specific components or parts of the system. Therefore localization assumes decomposition into subordinated functions.

What is exactly the starting point for decomposition? The first step is to understand which system is responsible for the phenomenon we are investigating. Bechtel and Richardson call this task *the identification of the locus of control*. This 'locus' is the 'place' where we think a certain phenomenon occurs. Identifying the right locus of control is a demanding task since it involves the identification of the right level of organization, the boundaries of the system itself and it has important consequences on how we decompose the system. For instance, in studying respiration it was far from being clear how this process actually occurred, until it was found out that the locus of control of respiration – where respiration happens – is the cell (Bechtel and Richardson 2010, Chapter 3 Section 4). The initial task of the identification of the locus of control implies the 'segmentation' of a system from its environment (see Figure 4.1).

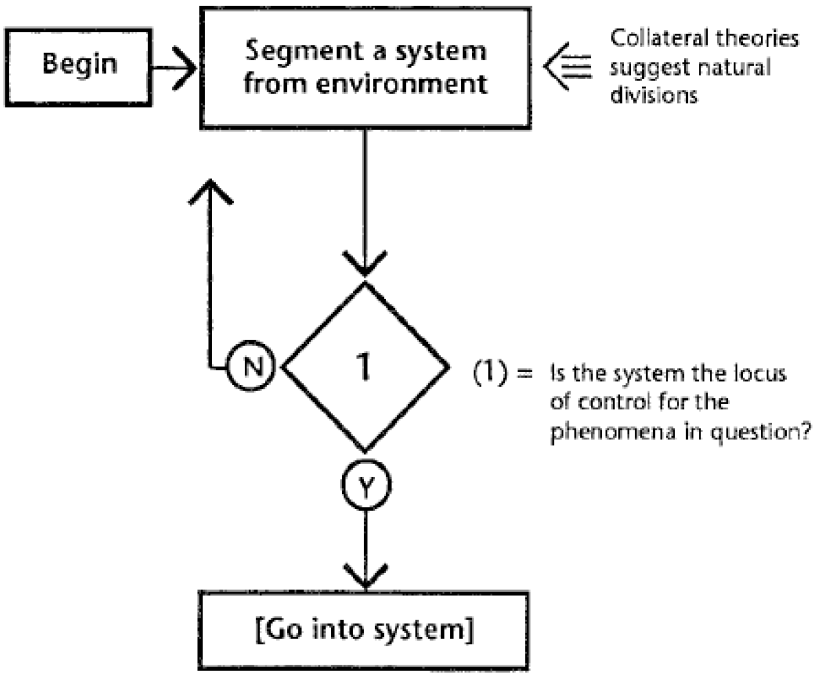

Figure 4.1 The identification of the locus of control. Taken from (Bechtel and Richardson 2010)

The next step then is to identify how the system performs what is required to perform. In other words, now the task is to go deeper into the mechanism underlying the behavior. This is provided by *direct or simple localization*. In particular, direct localization proposes a hypothesis of how the system can be decomposed into a set of components (Bechtel and Richardson 2010, p 63). Again, here there are contentious assumptions about the system, namely that it can be easily decomposed; that activities are not distributed among several parts, so that each activity can be assigned to one part; that by computing all parts' activities we will obtain the whole behavior we are investigating. Direct localization is similar to the task of identifying the locus of control, but with a substantial difference. What is different is the level of analysis. If identifying the locus of control requires the segmentation of a system from its environment, direct localization requires the segmentation of the system itself into components somehow relevant for the explanatory task.

Therefore, direct localization is placed within the system, and hence it is at a lower level of analysis than the level of determining the locus of control. Both determining the locus of control and direct localization are the results of the interplay between decomposition and localization. In the case of the locus of control, scientists decompose the activities of nature and localize some of them in a system, while in the case of direct localization we treat the locus of control as a set of components susceptible to the

decomposition strategy, and then we localize activities within components. This again can be shown with a diagram taken from Bechtel and Richardson's book (Figure 4.2).

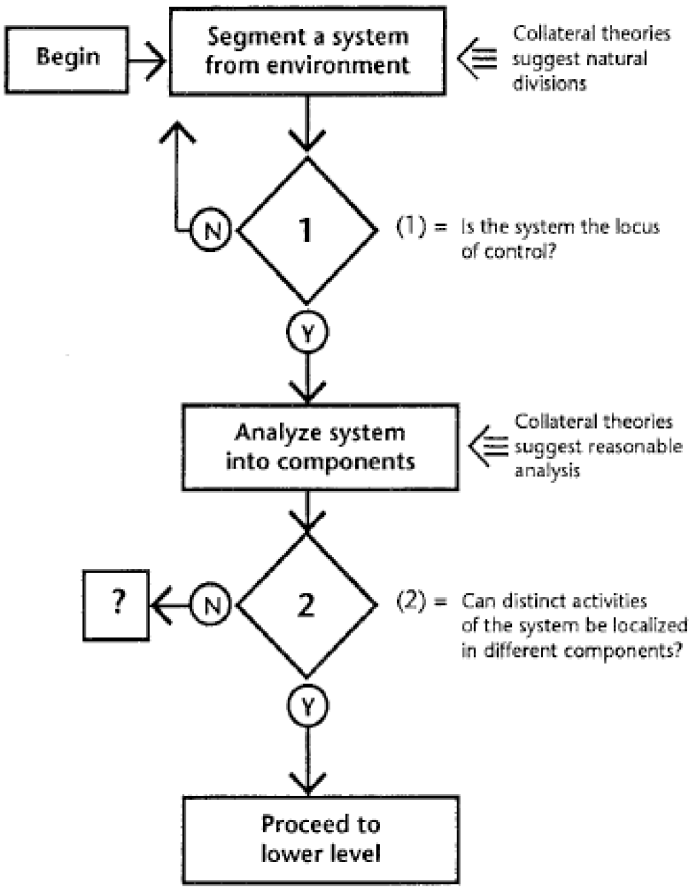

Figure 4.2. A diagram showing direct localization. Taken from (Bechtel and Richardson 2010)

While direct localization can provide strong evidence (for instance, by means of inhibitory or excitatory strategies) that a particular component is responsible for the behavior of a system, this is not enough to describe the exact mechanism explaining the phenomenon of interest. It localizes some components but it does not explain *how*. To explain how, we need a lower level of analysis. Without going too much into details, what is crucial here to point out is that again decomposition and localization come into play in going lower. In particular, localization has to *demonstrate* that entities or processes perform functions identified in decomposition, which is more than direct localization. Actually, direct localization is only - in Bechtel and Richardson's own words - an *insufficiently constrained localization*. This is important, as we will see in Chapter 5. As an example, at the time of first edition of Bechtel and Richardson's book (in 1993), some researchers claimed to have identified genetic markers on chromosome 11 of the human genome correlated to manic depression in the Amish community of Pennsylvania. This is a claim of direct localization. However, the evidence - Bechtel and Richardson continue - is purely

correlational and it gives no clue on whether certain genes are causal agents or simply markers.

A different case, again reported by Bechtel and Richardson (2010), is about Alzheimer's disease where localization and decomposition go deeper, by claiming that perhaps amyloid proteins are involved in the development of the disease. This evidence is stronger, since it suggests at least a possible mechanism. It should be emphasized that a good localization is one that suggests strategies to go deeper in another localization. But when exactly a mechanistic explanation is complete? In my humble opinion, Bechtel and Richardson do not suggest when the level of analysis is sufficient to provide a good mechanistic explanation. It seems that there is no ending in going lower and lower in the level of analysis, and that there are no good mechanistic explanations, but just some that are better than others. Darden's work partially can make sense of this problem (Machamer et al 2000; Darden 2006, Craver and Darden 2013; Darden and Craver 2002).

2.3 Strategies for discovering mechanisms (ii): schema instantiation, backward/forward chaining and anomaly revision

In an impressive number of works, Lindley Darden has investigated the strategies for discovering mechanisms (see for instance Darden 2006; Machamer et al 2000; Darden and Craver 2002; Craver and Darden 2013). According to Darden, there are four different stages in discovering mechanisms: characterizing the phenomenon, constructing, evaluating and revising a mechanism schema. As in Bechtel and Richardson's characterization[34], scientific discovery is here depicted as a piecemeal process, from an unconstraint sketch, to a more elaborated and constrained mechanistic description. 'Constraint' here is a keyword. Also in decomposition and localization, there are a number of constraints that limit the way the problem space can be decomposed and conceptualized.

According to Darden, when we start to discover mechanisms, how we characterize the phenomenon to be explained constrains the on-going mechanistic description. In other words, "characterizing the phenomenon prunes the space of possible mechanisms (because the mechanism must explain the phenomenon) and loosely guides the construction of this hypothesis space (because certain phenomena are suggestive of

[34] In Bechtel and Richardson's work, there are *at least* three kinds of constraints: (a) Phenomenological constraints, i.e. the way a phenomenon is characterized has clearly important consequences on the mechanistic descriptions that can be possibly elaborated; (b) operational constraints, i.e. available experimental techniques influence the way we conceptualize a decomposition task; (c) physical constraints, i.e. the background knowledge about the physical realization of lower-level components

possible mechanisms)" (Craver and Darden 2013, p 52). Characterizing a phenomenon means employing the language of a given field with all its explanatory concepts and the store of entities and activities of that field. All these resources then can be combined to elaborate mechanism scheme. For instance, in the case of cancer genomics described in Chapter 2, the phenomenon 'cancer' is characterized in terms of accumulation of somatic mutations and how the phenomenon is explained is by means of those somatic mutations and not, for example, by means of structural variations. The characterization of the phenomenon is the starting point of the discovery process. Consider a simple example. If one seeks to understand how two populations of cells communicate with one another, one might consider different scenarios (conceptualized from background knowledge of the field) such as an interaction among cell-surface proteins, paracrine mechanisms, etc. Therefore the starting point of the discovery process is populated by several possible mechanisms. The store of the field and the characterization of the phenomenon frame the discovery process. This is exactly a heuristic, in the sense of a way to attack the epistemic complexity of a problem space by limiting the possible configurations of the problem space itself.

Another important concept playing a key role here – especially in determining the quality of a mechanistic description – is the concept of *productive continuity*. As I have already mentioned, the quality of an explanation lies in filling the gaps in a causal chain connecting set-up and termination sides of the causal chain itself. The idea is that a good mechanistic description must show all the steps – with no gaps – of the production of a phenomenon[35]. With the characterization of the phenomenon and the concept of productive continuity as a gold standard, scientists can start the discovery of mechanisms. There are several strategies that can be used which are not exclusive.

Schema instantiation is the first strategy. One starts with a very abstract characterization of the mechanism, full of black boxes, and then she looks for components fitting placeholders. In this way, the mechanism schema is made less abstract by instantiating it. Where do abstract characterizations come from? One source is, again, previous knowledge in a field. One may hypothesize the type of mechanism involved and take a pre-existing framework as the starting point of her schema. There are many types of mechanisms that one may consider, such as transport mechanisms, control mechanisms, replication mechanisms etc. Darden (2006) makes the example of explaining some adaptations. In this case, we will look for a *selection schema*, thereby considering a stage of variant production, then a selective interaction challenging the variants, and finally a differential benefit for some of the variants. Other sources employed are *analogies*. One can draw 'analogies' from neighbouring fields. Alternatively,

[35] Clearly there is a problem here: Are we sure that we are able to include, in principle, all the entities contributing to the mechanism?

a source of analogies is the history of science. Once the schema is sketched, then one can start looking for entities and activities playing a specific role in the abstract schema. The idea is that one has to fill black boxes in a piecemeal fashion as soon as evidence for the components playing a specific role is found.

Another strategy guiding mechanism discovery is *modular subassembly*. This strategy requires more ingenuity than schema instantiation, since it is mainly reasoning about how mechanism components might be combined. The idea is that "[o]ne hypothesizes that a mechanism consists of (perhaps known) modules or type of modules. One cobbles together different modules to construct a hypothesized mechanism" (Darden 2006, p 286). It goes without saying, discovering 'modules' has important consequences on how future discovery strategies based on modular subassembly might be organized. One famous example of 'module' is *Pax6* and a group of related genes playing an important role in eye formation in invertebrates. Therefore, knowledge of a conserved module such as the one of *Pax6* will aid the construction of a plausible target mechanism of eye development in a new species (Craver and Darden 2013, p 75).

If schema instantiation provides the overall framework for a mechanism and modular subassembly proposes working components, the strategies of *backward and forward chaining* provide a finer grain analysis. These strategies enable scientists to reason about one part of the mechanism on the basis of other parts of the mechanism. Forward chaining uses earlier stages of mechanisms to reason on what might happen at a latter stage. Backward chaining is the opposite of forward chaining in the reasoning direction. In forward chaining one looks for 'activity-enabling' properties to reason forward in the productive continuity of the mechanism, while in backward chaining one looks for 'activity signatures' to reason backward. Craver and Darden (2013) provide the example of the finding of a hydrogen bond in a later stage of mechanism that somehow implies the existence of polar molecules with weak charges that have been neutralized in a prior stage of the mechanism.

Through the use of these strategies, mechanism sketches have to be evaluated and accordingly revised. Different types of anomalies can provide clues on how to evaluate and revise a hypothesized schema (Darden 2009). For instance, there are *compositional anomalies*, indicating that a certain entity, because of its composition, cannot accomplish what is taken to accomplish in the productive continuity of the mechanism. Another anomaly is the *temporal anomaly*: when something occurs more rapidly than expected by the sketch of the mechanism. Both anomalies indicate "the need for a change during the historical development of our understanding of the mechanism" (Darden 2009, p 50). Anomalies can be revealed by inhibitory and excitatory strategies, as "[p]oking and prodding an experimental system may reveal the presence (or absence)

of component entities and activities that then will need to find a place in a more articulated sketch" (Darden 2006, p 291).

To sum up, discovering mechanisms in Darden's perspective is a piecemeal and gradual process of refining incomplete sketches, and scientists fill gaps in sketches by consulting the store of a field, by combining modules, by reasoning backward/forward, by playing with experimental systems and by correcting anomalies. In the next section, we will see all these strategies in-the-making in the case study of the discovery of protein synthesis.

2.4 Using heuristics: the discovery of the mechanism of protein synthesis

The discovery of the mechanism of protein synthesis has been widely discussed by historians of science and philosophers of science. Darden and Craver (2002) use the history of this discovery as a perfect instantiation of the use of strategies for discovering mechanisms. There are three main reasons why this episode is a good case study. First, while it is not the case that *any* episode in discovering mechanisms in molecular biology should embed all the heuristic strategies so far identified, the discovery of the mechanism of protein synthesis seems to encompass *at least* three very important strategies (schema instantiation, forward/backward chaining and anomaly revision). In other words, it is a good example to show strategies of discovery. The second reason lies in the fact that it was a groundbreaking discovery, catalyzing many other discoveries in the 60s and 70s. Third, the discovery of the mechanism of protein synthesis is also particularly interesting for the history of molecular biology, since it is the result of the joint effort of two disciplines (genetics and biochemistry), it represents a remarkable strengthening of the shift from the protein paradigm to the nuclei acid paradigm, and it was radically informed and guided by the informational metaphor. For all these reasons, several commentators analyzed this episode (see for instance Burian 1996; Kay 2000; Morange 1998; Rheinberger 1997; Darden and Craver 2002).

To put it in very simple terms - *and to abstract from particulars I am not interested in* - the protein synthesis (or at least the formation of the primary structure of proteins) works as follows. The conceptual starting point of the mechanism is the DNA double helix. In the nucleus, one strand of the DNA double helix is used as a template to synthesize a pre-messenger RNA (pre-mRNA). Next, pre-mRNA is subjected to a chemical process called *splicing*, where some bases are removed (introns) from the template and the remaining are assembled together (exons) to form mRNA. At this point mRNA is said to include the *coding sequence*, which is a sequence of units of three nucleotides called *codons* as well as other untranslated regions at both ends of the transcript. The mRNA

then moves to the cytoplasm. In the cytoplasm a complex of macromolecules called ribosomes binds to the mRNA at what is called *the start codon*. The recognition of the start codon is accomplished by another RNA molecule called transfer RNA (tRNA). There are then others tRNAs which function as sort of adaptor molecules which actually link the triplets of mRNA to specific amino acids. In this way, tRNAs guide the association between codons and amino acids - thereby forming polypeptide sequences - until the stop codon. The outcome of this process is a sequence of assembled amino acids forming the primary structure of proteins.

The history of the discovery of such a mechanism (except for the part on splicing) is based on the efforts of two different groups of scientists. On the one hand, the biochemistry group of Paul Zamecnik was trying to grasp the mechanism for assembling polypeptides. On the other hand molecular biologists the likes of James Watson and Francis Crick were trying to understand the genetic code and how the order of DNA bases is related to the order of amino acids in proteins. Zamecnik described these efforts as the building of a tunnel, where "digging is going on from two sides of this mound of uncertainty, in the hope of meeting in the middle" (1962, p 47). The idea was somehow to connect the hereditary message of the gene and its expression by means of the role of RNA that was increasingly becoming important both for molecular biologists and biochemists.

The biochemistry group was reasoning from the end of the mechanism of protein synthesis[36]. It is a typical backward chaining "from peptide bonds to the mechanisms of polypeptide assembly, focusing on chemical reactions and energy requirements for such strong covalent bonds to form" (Darden and Craver 2002, p 5). They were focused on in-vitro studies, employing a sort of decomposition and localization interplay by identifying the locus of control in the cell, decomposing it into parts, and analyzing the function of its parts. Therefore they started with a specific store of entities and activities organized around chemical structures and reaction schemes of small molecules like peptides and covalent bonding reactions. They were trying to understand energetic intermediates between free amino acids and their linkage in polypeptides. Accordingly, they were filling black boxes of chemical mechanism scheme and equations, which were their starting points (i.e. instantiated scheme). The intermediate between free amino acids and their linkage turned out to be aminoacyl adenylate, but they also recognized macromolecular complexes in the cytoplasm made of RNA and proteins (that they called microsomes) where the polypeptide synthesis took place, as well as another soluble RNA lighter than microsomes' RNA. However, it was not entirely clear the role of RNA. Here there is not

[36] Clearly 'the end of the mechanism' has to be understood in an epistemic sense

only backward chaining, but also decomposition and localization. As a matter of fact, how the phenomenon was characterized - with known properties of polypeptides and in general the store of entities and activities - imposed important constraints on how the possible mechanism was understood, which entities they were expecting to fill black boxes, and how they were trying to localize them.

The molecular biologists were reasoning forward from DNA to ordered amino acids sequences in proteins by means of a scheme based on the informational metaphor. Watson and Crick transformed a biological problem into a problem of information theory, thereby thinking in terms of macromolecules and their informational content. The informational metaphor came from a suggestion of the astrophysicist George Gamow who suggested to Watson and Crick that, in order to understand how the order of DNA bases correspond to the order of amino acids in proteins, the problem did not need to be conceived in terms of biochemical reactions, but rather as a coding problem. Therefore they started reasoning forward from DNA by decomposing the problem of protein synthesis in terms of information-transfer activities, trying to localize the entities involved in this information transfer.

Biochemists and molecular biologists used different techniques and conceptualizations. Biochemists started with proteins, while molecular biologists started with DNA. Biochemists traced the flow of energy, while molecular biologists traced the 'information flow'. While biochemists were agnostic on the role of RNA, molecular biologists made several attempts to grasp its role into the phenomenon they were investigating. At first, since the structure of DNA was important in understanding many things about DNA, *analogously* they tried to determine the structure of RNA in the same way. It is a typical example of using neighboring fields by analogy to develop a mechanism schema. However, this attempt was unsuccessful. Then Crick elaborated the so-called 'adaptor-hypothesis'. According to this hypothesis there are 20 adaptors (one for each amino acids), each of which can specifically bind to a coding template of RNA. Zamecnik heard about this conjecture, and he linked it to the discovery of the smaller and soluble RNA, that was also found to bind to amino acids. This small soluble RNA came to be called *transfer RNA*.

With the characterization of the mechanism centered around tRNA, it was still necessary to grasp how information was transmitted from DNA to RNA templates. Pardee, Jacob and Monod discovered this by means of the so-called PaJaMo experiment. In *reasoning forward* Pardee, Jacob and Monod discovered a *temporal anomaly* in the *productive continuity* of an early mechanism schema of protein synthesis elaborated by Watson and Crick. They were investigating the insertion of the gene for the enzyme of B-galactosidase into a bacterium, and they observed that as soon as the gene entered the

bacterium lacking that gene, the enzyme started to be synthesized too early. At that time, scientists believed that such a synthesis would take much longer, because of the need to assemble microsomes from scratch. In particular, if the ribosomal RNA "had to be synthesized on the incoming DNA (the functional gene), and then the ribosomal particle had to be assembled, one would not expect the very rapid initiation of protein synthesis" (Darden and Craver 2002, p 15). Therefore, they hypothesized the existence of another type of RNA carrying information from DNA to microsomes. This turned out to be *the messenger RNA*. The story is not over, and I have neglected details of other groups working on the same mechanism, but for the purpose of showing how heuristic strategies work this is enough.

Let us now identify concepts and heuristics playing a key role in this story (Darden and Craver 2002). Backward and forward chaining, as well as decomposition and localization, have been already emphasized: Molecular biologists were reasoning forward to DNA, while biochemists were reasoning backward from polypeptides. *Productive continuity* has been characterized in two ways. Molecular biologists proposed productive continuity in terms of information transfer (preservation of linear order from DNA bases to amino acids passing through mRNA), while biochemists conceptualized productive continuity in terms of flow of matter and energy represented in terms of chemical equations. These two senses were then integrated in the end. *Schema instantiation* took place in two ways (see Darden and Craver 2002, pp 20-21 for the details of the scheme). Biochemists start with schemes taken from chemical reactions and equations of the form of

1. Amino acids + ATP = other centrifuge fractions ➔ activated aa complex ➔ protein

while molecular biologists start in terms of information flow:

2. DNA ➔ helical RNA ➔ protein

And they tried to fill black boxes in these simple schemes accordingly.

To sum up, in this episode we find most of the heuristics that have been discussed. Decomposition and localization, the store of fields and analogies played a key role in setting the starting points of both molecular biologists and biochemists, while schema instantiation and backward/forward chaining as well as anomaly revisions were of paramount importance to develop the final mechanism description. Now that I have explained strategies of mechanism discovery in molecular biology, I can turn to molecular oncology, and try to see the relation between Darden's (or Bechtel and Richardson's) discovery strategies and my framework of discovery in order to compare 'traditional' molecular biology to data-driven molecular biology.

3. TRADITIONAL MOLECULAR ONCOLOGY

In this section, I analyze what I call 'traditional' molecular oncology from three different angles. First, I will describe a case study (Guo et al 2012) taken from Weinberg's lab following the narrative of scientific experiments as it appears in the original article. Next, I identify Bechtel and Richardson as well as Darden's discovery strategies in the case study. Finally, I reconsider these strategies and the case studies in their relations to the tripartite framework depicted in Chapter 1, considering what happens within each phase of discovery, the inferences used and the epistemic values guiding the whole process. But before starting, it is worth to spend few words on the history of molecular oncology.

3.1 Complexity, simplicity and complexity again: The strange case of molecular oncology

Molecular oncology developed around the oncogene paradigm (Morange 1998, Chapter 19). The stabilization of this paradigm took place between 1975 and 1985 and it was based on the idea that cancer, whatever it is exactly, "is the result of the activation, by modification of over-expression, of a highly conserved family of genes called oncogenes" (Morange 1998, p 219). An oncogene is a gene that, when overexpressed, has a direct or prominent role in tumor initiation or development. Oncogenes are usually coupled with tumor suppressor genes, which are genes that prevent cells from entering in a neoplastic state. In my understanding, when we talk about 'the oncogene paradigm' both oncogenes and tumor suppressor genes are included. In Chapter 19 of Morange's book (1998) one may find the historical details about the decade between 1975 and 1985 but I think that, in order to introduce molecular oncology, Weinberg's essay (2014) *Coming Full Circle – From Endless Complexity to Simplicity and Back Again* is a more tantalizing source for at least two reasons. First, since Weinberg is essential in the controversy motivating this work, and since the case study of traditional molecular oncology I am going to employ is from Weinberg's lab, introducing this field through his eyes seems more coherent with my plans. Next, Weinberg does not focus only on historical details, but rather on how the complexity of cancer has been epistemically attacked.

In his essay appeared on *Cell* (2014), Weinberg traces the history of molecular oncology from the point of view of the emerging generation of scientists in the 1970s. At that time, the mechanisms of cancer development were completely unknown, and senior professors seemed to be fine with that. The younger generations of biologists, of which Weinberg was part of, endorsed a reductionist approach to the biological problem, claiming that the biology of cancer could have been understood in terms of few molecular mechanisms. Older professors disagreed, claiming "that cancer was really much too

complicated to be understood through simple molecular mechanisms. Indeed, they portrayed our reductionism as simplistic if not simple-minded" (Weinberg 2014, p 267). The real history of the rise of the oncogene paradigm dates back exactly at 1971, when President Nixon's War on Cancer was declared, and an impressive flow of money were redirected to cancer research. To tell the truth, this 'war' was strictly rooted on the conviction that cancer was ultimately a disease of viruses infecting the genome. But by the mid-1970s the search for human retroviruses developing cancer turned out to be rather unsuccessful. Another idea was increasingly gaining recognition, namely that cancer was a disease of identifiable genes somehow able to overrule the far vaster genome of normal cells pushed into neoplastic states. The so-called Varmus-Bishop discovery (Stelhein et al 1976) of the *src* proto-oncogene was a step towards the stabilization of this idea. This discovery showed that, in rodents, normal cells carrying a specific gene that could be 'kidnapped' and modified by a retrovirus, could be transformed into cells in a neoplastic state. Moreover, Weinberg and his fellows were 'energized' (his own word) by some works of Bruce Ames correlating carcinogenic potency of a chemical species to its mutagenic activity. This suggested that cancer cells could be conceived as mutants, and that mutant genes were capable of driving malignant cell proliferation. Here came the idea that the birth of a malignant cell could be the result of few molecular events. Therefore in the mid-1970s, everyone started to improve the ability to introduce transforming genes (oncogenes) via transfection into normal cells.

The birth of molecular oncology shows some of the strategies depicted in the previous sections. First, it is shown that in the 1970s people were looking for a way to attack the complexity of cancer. People first looked for *a locus of control*, identified with the genome. Then, people tried to do a *direct localization*, first focusing on retroviruses, and then establishing a correlation between specific genes and malignant cell states. The problem "how do we explain cancer?" was *decomposed into a problem* of localization of few genes responsible for tumor development. From this conjecture, the development of molecular oncology is just a logical consequence, namely the effort of going deeper into the mechanistic details of how specific genes were actually forcing normal cells to enter neoplastic states.

In 1982, Weinberg's group and others (Shih and Weinberg 1982; Pulciani et al 1982; Goldfarb et al 1982) isolated through cloning a transfected human bladder carcinoma gene. This human oncogene was a homolog of the *Ras* oncogene discovered by tumor virologists. By 1982, some results fueled two important (and misleading) convictions guiding decomposition and localization in molecular oncological research. First that in order to transform normal cell into cancer cells, a single specific gene was enough. Therefore, the decomposition of the phenomenon 'cancer' was (conceptually) based on the *direct localization* of a key entity – *a specific gene*. *Instantiating a schema*, and then

reasoning *forward and backward* were thought as being relatively easy task if based on just one entity – *a specific gene*. Next, this discovery was misleading also because the bladder carcinoma oncogene differed from its normal counterpart by *only* a single point mutation. Therefore cognitive strategies put in place to elaborate mechanistic descriptions looked much easier since they could have as starting point just a single point mutation.

In the next few years, several articles undermined these convictions. In 1983 it was discovered that two mutant genes – collaborating with one another – were responsible for tumor initiation. But this was for rodents. The corresponding human cells required five distinct mutant genes.

Next, it was time for a debate about oncogenes and tumor suppressors: Which one is really responsible for cancer? It was soon revealed that both are equally important in tumor development. Vogelstein's pioneering work on colorectal cancer (Vogelstein et al 1989; Fearon and Vogelstein 1990) showed that the more colorectal tumors progressed, the larger the number of somatic mutations affecting both oncogenes and tumor suppressors was. The 1980s and the 1990s were years devoted to show how complex the phenomenon of cancer development was, and the repertoire of oncogenes and tumor suppressors grew exponentially.

Despite the emerging complexity of cancer, Hanahan and Weinberg tried to reason both analytically and synthetically to grasp (or to idealize) some sort of order beneath the chaos of cancer. In particular, they tried to identify some commonalities, namely some phenotypes common to all neoplastic states, which can be used as heuristics to analyze and to categorize tumors. In a famous critical review (Hanahan and Weinberg 2000) six hallmarks of cancer were identified, and in 2011 two more were added (Hanahan and Weinberg 2011) as well as two enabling capabilities of the hallmarks themselves. Figure 4.3, adapted from Hanahan and Weinberg (2011), provides a visual summary.

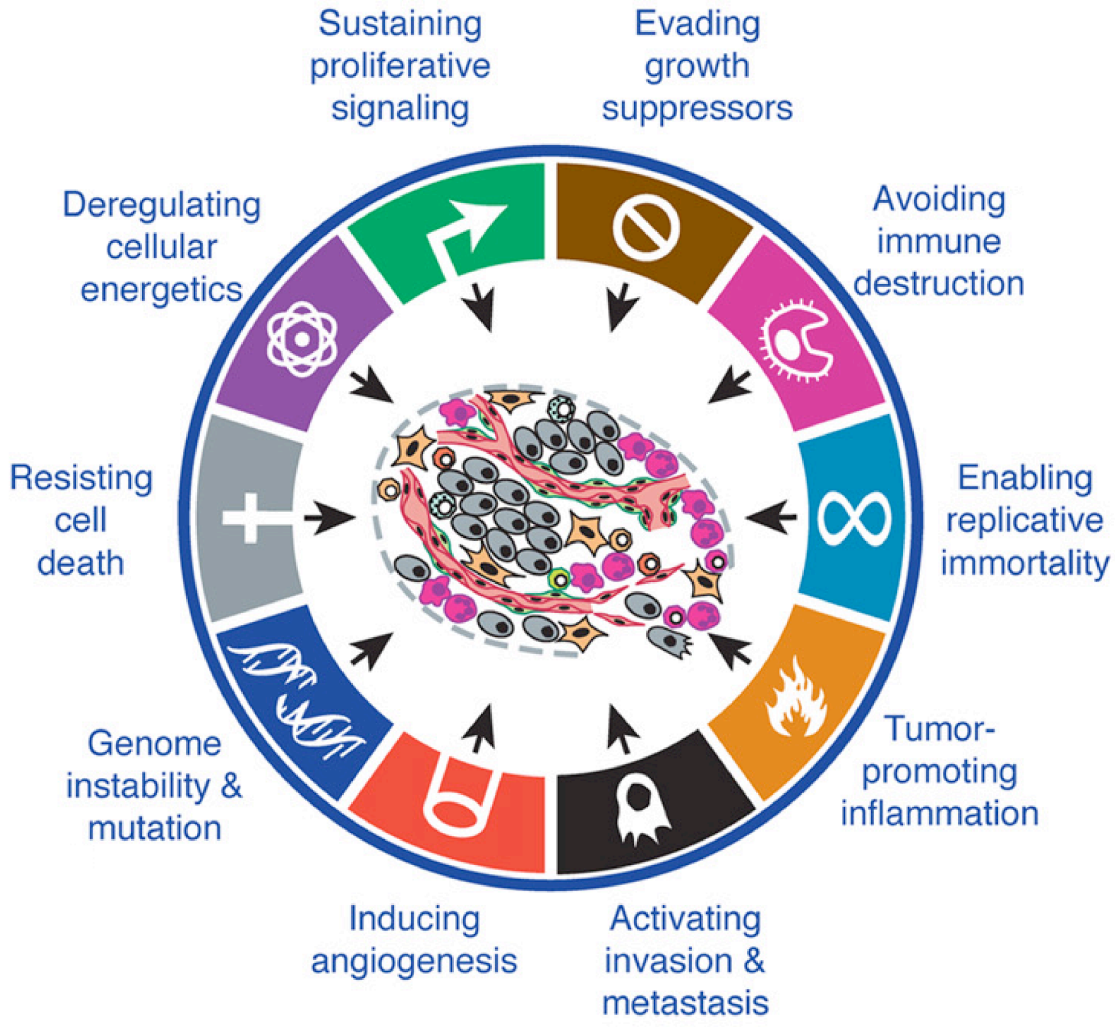

Figure 4.3. The hallmarks of cancer. Taken from (Hanahan and Weinberg 2011). The figure has been modified

It is very important to stress that the type of categorization provided by Hanahan and Weinberg provides a series of schemes or modules that can be used both for schema instantiation and modular subassembly, or that can be used to reason forward and backward from any point in mechanism description. Actually, for each hallmarks there are genes likely to be responsible for tumor initiation, types of mechanisms put in place or pathways to explore. Therefore reconstructing mechanisms of cancer can take advantage of such 'store' of entities, activities, modules and types of mechanisms. While Hanahan and Weinberg's work does not aim at *explaining* cancer complexity, it is nonetheless useful to attack the epistemic complexity as it is usually done in the tradition of discovering mechanisms.

In this section I have introduced the history of molecular oncology from the point of view of Weinberg, and his concerns about how we can attack the epistemic complexity of elaborating mechanism descriptions of cancer behavior. As I hope it is clear, molecular oncology practitioners tend to identify the locus of control in the genome, they localize

relevant parts in genes and mutations, they instantiate mechanism schemes and they try to reason forward and backward with the use of modules. In the next section, I discuss a case study taken from Weinberg's lab (Guo et al 2012).

3.2. Traditional Molecular oncology in the making

In (Guo et al 2012) Weinberg's group tries to get insights into the mechanism of induction and maintenance of adult stem cells and they find out that the mechanism they depict promotes tumorigenic and metastasis-seeding abilities in human breast cancer cells.

Adult stem cells are undifferentiated cells residing with differentiated cells in various tissues or organs. These cells can differentiate into the major types of cells of tissues and organs where they are located. They are aimed at repairing and maintaining the tissue or organ where they reside.

Weinberg's group found out that two transcription factors (TFs) – Slug and Sox9 – act cooperatively to induce the adult stem cell state in mammary stem cell by activating distinct autoregulatory gene expression programs. Moreover, they showed that co-expression of the two TFs promotes some typical cancer's hallmarks in human breast cancer cells. Actually in this article Weinberg's group showed much more, but these two results are enough. This is because here I am not (only) interested in the final results, but rather in how they set their starting point, and how they arrived at their conclusions about the cooperation of Slug and Sox9 in determining mammary stem cell state and human breast cancer. Here are the essential conceptual and experimental steps of this study.

1. Weinberg's group drew from two previous set of evidence the aims and the methods of their study. First, previous studies have shown how master TFs play key roles in determining cellular states, including stem cell states. However, for adult stem cell such evidence was missing. Therefore Weinberg's group decided to go deeper into this issue, in particular for epithelial tissues. They chose mammary gland as experimental system because this contains a small subpopulation of cells with robust stem cell activity. This means that the mammary gland (in murine) was considered as a powerful system to understand the mechanism they wanted to uncover, as well as providing stringent test for stemness. Next, Weinberg's group and others showed in previous studies that there is a surprising connection between epithelial-mesenchymal transition (EMT) program and the mammary stem cell. The idea was that the passage of both normal and neoplastic mammary

epithelial cells through EMT confers on the resulting cells some important features shared with mammary stem cell. In other words, EMT program drives mammary epithelial cells into mammary stem cells. Therefore, here Weinberg's group sought to understand the genetic pathway cooperating with the EMT program enabling such transformations.

2. Weinberg and his collaborators used different populations – sorted accordingly – of primary murine mammary epithelial cells. They measured the expression of mRNAs encoding ten TFs previously described as being implicated in EMT program and they find out that only Slug was significantly expressed in the mammary stem cells enriched population. Therefore, Weinberg's group started to analyze the idea of Slug as being implicated in the maintenance and induction of adult stem cell state. They expressed Slug into mammary epithelial cell to investigate the role of this TF in inducing mammary stem cell state. They did this by various technical means those details are not of interest here.
3. They found out that Slug was not able to induce the formation of mammary stem cells from differentiated luminal cells. Therefore they hypothesized that probably Slug requires another TF to do this. To identify the cooperating factor(s), they selected eight TFs known to play various roles in either embryonic or adult stem cell biology, and are also known to cooperate with Slug in early developmental processes. They discovered that Sox9 is likely to cooperate with Slug in forming organoids[37]. If Slug and Sox9 are implicated in transforming differentiated luminal cells into organoids, this means that they could be implicated in the transformation of luminal cells into adult stem cells which in turn can also give birth to other cell types. Therefore, they took this as preliminary evidence that Slug and Sox9 together can *at least* induce mammary stem cell state.
4. Weinberg and his colleagues expressed Slug and Sox9 together to see whether they convert differentiated luminal cells (as well as other differentiated epithelial cells) into mammary stem cells. Moreover, they knocked down either Slug or Sox9 to check whether the continued coexpression of the two TFs is required to maintain mammary stem cells. They were successful.
5. Since the cooperation of the two TFs was shown to be required for induction and maintenance of mammary stem cells state, Weinberg and his collaborators sought to understand how exactly this happens. First they found evidence that Slug and Sox9 activate distinct – but complementary – biological programs. They discovered that in differentiated luminal cells Slug upregulates the expression of mRNAs encoding five of six basal cell-associated TFs, while forced Sox9 expression of genes associated with luminal progenitors. In other words, Slug and Sox9

[37] Organoids are multicellular complexes containing various cell types

upregulate the gene expression of both basal cells and luminal progenitors. Therefore, while Slug regulates the basal lineages program, Sox9 has a key role in the luminal program and these two programs are known to be required for mammary epithelial cells to enter and reside in the mammary stem cells state.

6. They also found out that, when ectopically expressed Slug and Sox9 are turned off after some days, the gene expression programs they promote remain active. This is because ectopically expressed Slug and Sox9 also induce the expression of their paralogs promoting the stabilization of an autoregulatory network that contributes to the maintenance of the stem state.
7. The final mechanism, as far as I understand, is as follows. Slug and Sox9 in differentiated luminal cells activate two typical gene expression programs typical of basal cells and luminal progenitors which in turn promote the entrance and the maintenance of the typical mammary stem cell state. After some time, Slug and Sox9 also activate paralogs promoting an autoregulatory network which is able to maintain the mammary stem cell state.
8. Finally, Weinberg's group seeked to understand whether a similar mechanism act also in human breast cancer stem cells and they find evidence that Sox9 and, in part, Slug are required for maintaining robust tumorigenicity.

Let us now try to identify, step by step, some of the strategies discussed in previous sections within this case study.

1. In this step, Weinberg identified the locus of control and he set up a typical decomposition strategy coupled with schema instantiation. The *locus of control* here is the experimental system, in the sense that in order to understand the maintenance of adult stem cell, he chooses the mammary gland for all the reasons previously listed. *Decomposition* comes along the fact that, in order to study the determination of the mammary stem cell state, Weinberg decomposed the whole activity into a particular set of sub-activities: The activities pertaining gene regulation. At the same time, here comes *schema instantiation*, as the idea of focusing on TFs regulatory activities is coupled with a whole literature having some sort of *'ready-made' modules*, entities and additional activities to be considered when dealing with such regulatory activities. Moreover, the previously shown connection between the EMT program and mammary stem cell is *a precious source of knowledge* to set up an initial set of possible mechanisms governing the phenomenon of interest.
2. This is a phase of localization and forward chaining. *Localization* is done basically in two ways. First, from the literature Weinberg's group considered a set of TFs

that could possibly play a role in the induction and the maintenance of the stem state. Then, a series of inhibitor and excitatory strategies were put in place to localize which of these TFs is likely to play a key role. When they focused on Slug, they started *reasoning forward* playing with *inhibitory and excitatory strategies* again to get insights on its role.

3. In this phase anomaly revision and localization played a key role. *Anomaly revision* is important, since the hypotheses related to Slug did not seem to be consistent with experimental results. Here there is a *compositional anomaly* forcing Weinberg and his collaborators to modify their initial hypothesis. They then speculated that Slug could cooperate with one or more TFs, and they sought to localize them. By *drawing again from the literature*, they selected a set of other TFs and they discovered that Sox9 could be Slug's 'collaborator'.
4. Here *localization* by means of *forward chaining* is fundamental. Accordingly, from the result of phase 3 they reasoned forward speculating the effects of the cooperation of Slug and Sox9 and they verified whether their localization hypothesis was corroborated by experimental results.
5. Here again is a matter of *forward reasoning*. After having clearly *localized* Slug and Sox9 as key regulators, they sought to fill black boxes (understood as genetic pathways) in their *flow of reasoning*. They put in place a typical *forward chaining* by reasoning on the possible genetic programs activated by Slug and Sox9 and they found out that these genetic programs while being strictly distinct, they are complementary.
6. Next, they considered another *anomaly*, namely the fact that sometimes while ectopic expression of Slug and Sox9 is turned off, the gene expression programs they promote remain active. Therefore, they slightly modified their hypothesis according to the observations that Slug and Sox9 activate a robust autoregulatory network that maintains active the gene expression programs.
7. They finally came out with a mechanism description able to fill (most of) black boxes considered in schema instantiation, and that seemed to make sense of *productive continuity*.
8. They hypothesized also that a similar mechanism can be active also in human breast cancer stem cells. Here *drawing analogies from a close field* is particularly relevant.

3.3 The tripartite framework and epistemic values in traditional molecular biology

Let us now see how this picture is related to the tripartite framework elaborated in Chapter 1. The first phase is hypothesis generation, i.e. the generation of one or more hypotheses about a certain phenomenon or state of affairs. These hypotheses are in a very abstract form and they are usually created by means of a combination of intuition, creativity, analogies, knowledge of a certain field, background assumptions etc. In the second phase (named weak assessment, prior assessment, theory pursuit, etc) adding new evidence develops hypotheses. Here there is a twofold movement: On the one hand hypotheses that turned out to be false or less probable are discarded, while more promising hypotheses are further developed, making them less abstract. Here eliminative inferences and retroduction are likely to play a key role. In the third phase (called 'justificatory phase'), confirmation in all its forms is prominent, and final hypotheses are subjected to more stringent tests. Now, step by step, let us how these three phases can be identified in Weinberg's case study, and which inference are used.

1. In this step, Weinberg's group set up some of the *background assumptions* that are fundamental in order to generate an initial set of hypotheses about the phenomenon under scrutiny. First, they assumed that the aspects of the phenomenon they wanted to explain depend on TFs regulatory activity. Moreover, hypotheses depended on the previous knowledge about the connection between the EMT program and mammary stem cells.
2. Here there is *hypotheses generation*, as well as *prior assessment*. Weinberg's group generated a set of hypotheses drawn from the literature about TFs regulation in induction and maintenance of the stem state. Hypotheses are of the form of 'such TF regulates mammary stem cell state'. Next, it started prior assessment. Weinberg and his collaborators provided evidence that only one of the hypotheses members of the initial universe of conjectures was promising, namely the hypothesis that Slug can play a role in the phenomenon of interest.
3. This is a fundamental turning point in *prior assessment*. The hypothesis 'Slug regulates mammary stem cell state' is developed, by adding new evidence, into 'Slug and Sox9 cooperate to regulate mammary stem cell induction and maintenance'.
4. Similarly to step 3, Weinberg and collaborators collected evidence to develop the primary hypothesis.
5. Again this is a phase of *prior assessment*, with a push towards the mechanistic details of the hypothesis, now becoming 'Slug and Sox9 in differentiated luminal

cells activate two typical gene expression programs typical of basal cells and luminal progenitors which in turn promote the entrance and the maintenance of the typical mammary stem cell state'.

6. *Final formulation of the hypothesis*, namely 'Slug and Sox9 in differentiated luminal cells activate two typical gene expression programs typical of basal cells and luminal progenitors which in turn promote the entrance and the maintenance of the typical mammary stem cell state'. Here more stringent evidence for the whole hypothesis was added (*justificatory phase*). Moreover, another hypothesis is added, namely that 'Ectopically expressed Slug and Sox9 activate paralogs promoting an autoregulatory network which is able to maintain the mammary stem cell state when they are turned off'.
7. Here more stringent evidence for the whole hypothesis was added (*justificatory phase*)
8. The hypothesis is transferred into a different field, and subjected to prior assessment and justification

Therefore, it seems that hypothesis generation is guided by a heterogeneous set of background assumptions. These include *previous knowledge in a specific field* (in the case of TFs regulation), as well as *analogies from close fields* (as in the case of the final mechanism transferred to breast cancer). Prior assessment is driven mostly by *eliminative inferences* (when TFs are eliminated from the initial set of hypotheses), but it should be pointed out that also *retroduction* plays a prominent role. For instance, when the hypothesis about Slug seems to be undermined, Weinberg reasoned that their previous evidence could actually make sense only by postulating the existence of a partner in regulation. This is a striking case of abduction/retroduction. On the other hand here the justificatory phase – and how the final mechanism is evaluated – makes use of *methods and strategies which are not different in principle from the ones of hypothesis development*. Therefore one might say that here there is no difference between weak and strong evaluation. However, I would say that differences between weak and strong evaluation in such intertwined mechanistic constructions are fuzzier than in data-driven biology, where hypotheses development and justification are strictly separated. This is not to say that there is vagueness everywhere. Actually, there is a sense in which mechanism descriptions are strongly evaluated. Once a molecular biologist has elaborated a rich mechanistic description, the fact that no further anomalies emerge from experiments (and hence evidence that can undermine productive continuity) can be taken as a test that the mechanism description is sufficiently detailed and complete. This is a sort of final hypothesis testing, something that can close a 'scientific story' or a narrative. But we do not know that we are doing such a strong evaluation until we see that

productive continuity is met and no anomalies emerge. Therefore, we recognize strong hypothesis testing only *retrospectively*. Therefore, if one makes some technical mistakes in doing such experiments and further anomalies emerge, then this final set of experiments would not be recognized as a strong hypothesis testing, while in data-driven biology *we are always aware that we are in the justificatory phase*. In other words, justificatory phase in traditional molecular biology and data-driven biology are quite different because in traditional molecular biology the justificatory phase is recognized only retrospectively. Moreover, there is also another retrospective way in which a mechanism description can be put at test in a strong sense. A good mechanism description should also make sense of the observations done to construct the description itself.

It is easy to show how in general strategies identified by Bechtel, Richardson and Darden have counterparts in my tripartite framework. For example, *hypothesis generation* can be thought as the initial stage of *decomposition*, when a phenomenon is decomposed into subset of activities. Decomposition implicitly generates also a list of possible entities that can be responsible for the activities, usually drawing from existing literature. This is also the case of the *constraints imposed by the nature of the phenomenon* under investigation. How we understand the phenomenon comes along with a set of possible hypotheses about the organization of the phenomenon itself. Therefore decomposition and how we understand the nature of the phenomenon correspond, roughly, to the phase of hypothesis generation. *Localization strategies*, as well as *forward/backward chaining and anomaly revisions*, are all ways of 'weakly' evaluating an initial hypothesis, as well as means of developing conjectures. Finally, as I have just said, when in an experiment (or a set of experiments) no further anomalies emerge and productive continuity seems to be met, mechanism descriptions *are taken to be strongly evaluated*, but only retrospectively.

Now let us move to the identification of epistemic values in traditional molecular biology. Epistemic values are criteria influencing hypothesis generation, prior assessment and theory choice. Moreover, they establish those criteria that fill the content of scientific inferences, which are empty and somehow formal. Questions that can be answered by invoking epistemic values include: Why should scientists select a set of hypotheses instead of another? According to which criteria some hypotheses are worth to be developed? And in which direction should we develop hypotheses? Which are the features of a hypothesis that makes it an acceptable hypothesis? Now I identify the main epistemic values playing a role in traditional molecular biology.

In the phase of *hypothesis generation* the value of external consistency is particularly relevant. The way we choose a schema to instantiate, the consultation of the 'store' of the field, the use of 'ready-made' modules to build a sketch of the mechanism are all based on the idea that we should base our research to previous knowledge, that entire disciplines are the background assumptions we should employ as a starting point

and that, eventually, we must be consistent with this enormous corpus of knowledge in order to generate hypotheses. In the phase of prior assessment both external consistency and empirical adequacy are prominent. As a matter of fact, *to localize* entities – and therefore to eliminate or develop hypotheses for example about the role of TFs – one needs clearly to compare her hypotheses to observations. If hypotheses are not empirically adequate, they have to be modified accordingly or *tout court* eliminated. But there is also interplay between empirical adequacy and external consistency. In the case of Slug for example, the main hypothesis is modified and developed not only with observations that Slug alone is not sufficient to maintain stem cell state, but also by exploring literature on Slug and other TFs, thereby playing the card of external consistency to modify the hypothesis. In the justificatory phase, empirical adequacy also is prominent, since the final mechanism is weighted against (a) previous data generated during the construction of the mechanism description and (b) new data generated for testing the mechanism itself. Internal coherence, as in data-driven biology, is ubiquitous.

4. CONCLUSION

In this chapter, I have analyzed the structure of discovery of what I call 'traditional molecular biology'. First, I have introduced the field of molecular biology. Next, I have described the received view on discovery in molecular biology – based on the heuristic strategies to construct mechanism descriptions – and I show how they have been applied by philosophers to the history of the discovery of the mechanism of protein synthesis. Then, I have turned to molecular oncology. I have introduced the field by means of Weinberg's analysis of the reductionist saga of the oncogene paradigm. I have also analyzed a case study of molecular oncology 'in the making' taken from the work of Weinberg's lab. I have then developed a twofold analysis. On the one hand, I have shown how this case study can be embedded in the typical discovery strategies elucidated by mechanistic philosophers. On the other hand, I have shown how this case study can be also embedded in the tripartite framework elaborated in Chapter 1. Several inferences are put in place. In particular, hypothesis generation in molecular biology seems to be rooted in the use of analogies and derivation from background assumptions while prior assessment is based on an eliminative inferences framework. However, as I have noted, also retroduction can play an important role. In the justificatory phase, though more difficult to be identified than in data-driven molecular biology, confirmation is prominent. Finally, I have identified the epistemic values guiding inferences within each phase of discovery. In the next chapter, I shall finally compare data-driven and traditional molecular biology.

CHAPTER 5

A COMPARISON BETWEEN 'TRADITIONAL' AND DATA-DRIVEN MOLECULAR BIOLOGY

Chapter abstract

In this chapter I will compare the discovery strategies of traditional and data-driven molecular biology in light of the conceptual tools presented in Chapter 1. A preliminary result of this comparison will be that at the level of inferences and of *traditional* epistemic values the disagreement between traditional and data-driven biologists seems to be misleading. I will show that data-driven molecular biology is just a more regimented version of the strategies of discovery of traditional molecular biology. Still, the typical models elaborated by the two communities are substantially different as they value different *desiderata*. For this reason, I will go back to the analysis of epistemic values, and identify a particular category of values, which I shall call *quasi-epistemic*. I will then show how traditional and data-driven biologists actually endorse two different quasi-epistemic values (depth and generality, respectively) and, in the conclusion, I will argue that this difference can explain the disagreement between the two traditions.

1. INTRODUCTION

In the previous chapters I analyzed the structure of discovery of traditional (Chapter 4) and data-driven (Chapter 2) molecular biology by means of a tripartite framework for discovery developed in Chapter 1. I am now in the position of comparing the two approaches and to understand at which level the controversy motivating the present work lies. *Repetita iuvant*: There is a lively debate opposing two factions. On the one hand, there are scientists such as Robert Weinberg, Bruce Alberts and Sydney Brenner claiming that 'hypothesis-free' and sequencing technologies-based methodologies are not properly scientific and they are not effective and successful as traditional 'hypothesis-driven' approaches. On the other hand, practitioners as Todd Golub, Bert Vogelstein and Eric Lander argue that a data-driven and genomic approach is not only genuinely scientific, but it fosters drug discovery innovation by revealing biological insights that more traditional approaches would fail to uncover. There are clearly two different (though intertwined) aspects of the debate. On the one hand there is an epistemological problem, namely that there are two different approaches to molecular biology, and that these approaches generate two different kinds of knowledge, one of which is *in principle* inferior

to the other independently of the outcomes. On the other hand, there are empirical claims, e.g., that traditional molecular biology *is more successful* (in terms of scientific results) than data-driven molecular biology. Here I am going to analyze just the epistemological side of the controversy. However, I would like also to list few facts showing that – in molecular oncology – the cancer genome approach has generated insights that 'traditional' approaches did not provide.

There are some fundamental insights that came out directly of the massive projects of molecular oncology such as TCGA (Garraway and Lander 2013; Vogelstein et al 2013). First, there is the discovery that recurrent mutations in *IDH1/2* could link genetics and cancer metabolism (Figueroa et al 2010; Lu et al 2012). A second fundamental insight is that mutations can disrupt chromatin remodeling and DNA methylation in many cancers (Dolnik et al 2012). Next, it has been found out that mutations disrupting RNA splicing occur in many cancer types (Ellis et al 2012; Biankin et al 2012). Other insights come from studying particular kinds of tumors. For instance, many mutations disrupt Notch signaling in head and neck cancer (Stransky et al 2011), or mutations can deregulate squamous differentiation in lung squamous cancer (TCGA 2012). Moreover, cancer genomics screenings fostered the development of models of tumor evolution like punctuated progression as well as single catastrophic structural events. It is important to stress that all these insights came out of cancer genomics screenings that have the same discovery structure depicted in Chapter 2 and that many of these insights have to do especially with *big numbers*, meaning that these studies reveal trends across many types of cancer. Now the question is straightforward: Is this kind of knowledge – generated in that particular way - *inferior in principle* with respect to the knowledge that traditional approaches can possibly generate? But this question presupposes another more fundamental question: What is the difference between traditional and data-driven biology that make the approaches mutually irreducible? To answer these questions, first I make a step-by-step comparison of inferences and epistemic values within each phase of discovery of the two approaches. Next, I draw some conclusions on my analysis, arguing that the two approaches are not, at first glance, irreducible and that data-driven biology is a more regimented and standardized version of traditional molecular biology. Finally, I will identify the real reason for disagreement in the endorsement of a special category of epistemic values.

2. A COMPARISON OF DATA-DRIVEN AND 'TRADITIONAL' MOLECULAR BIOLOGY

In this section I compare data-driven and traditional molecular biology. There are two levels where the comparison takes place. First, there is the level of inferences used within each phase of discovery. Second there is the level of traditional epistemic values endorsed within each phase of discovery. Table 1 compares inferences used in data-driven and traditional molecular biology, while Table 2 focuses on epistemic values. Let us start with Table 1.

	Hypothesis generation	**Prior assesment and development**	**Justification**
Traditional Molecular Biology	Derivation from knowledge of the field Analogy	Eliminative inferences Retroduction	Confirmation
Data-driven Molecular Biology	Derivation from knowledge of the field	Eliminative inferences	Confirmation

Table 5.1. A comparison of the use of inferences in data-driven and traditional molecular biology within each phase of discovery

Hypothesis generation in data-driven and traditional molecular biology shares important features. Hypotheses are not generated, say, 'out of the blue' but they are mostly derived from the knowledge of the field. *For data-driven biology* it is relatively easy to show this. GWASs are based on the hypothesis that SNPs can have an effect on the development of common diseases including some types of cancer (e.g. prostate

cancer). From this bit of knowledge, GWASs practitioners can hypothesize that the SNPs detected by SNP arrays might be implicated in the development of the disease they are studying. Clearly, the generation of hypotheses does not entirely depend on background knowledge. It is unreasonable to think that, even before starting a GWAS, an epidemiologist can actually think that all the SNPs so far discovered can have a role in the disease. Actually, the universe of hypotheses is formed just after SNPs are detected. However, it is worth stressing that previous knowledge about SNPs actually *sets out the very possibility that a type of study such as a GWAS might be even conceived*. Therefore, in a sense we might argue that the existence of the hypothesis generation done in such and such a way in GWASs depends prominently on previous knowledge. In cancer genomics the situation is similar. The very fact that we endeavor to form a universe of hypotheses just after initial phases of sequences having certain features (e.g. we focus on somatic mutations) depends mostly on the previous knowledge accumulated claiming that cancer is (in many cases) the result of the accumulation of somatic mutations. Without this piece of knowledge, there will not be any initial universe of hypotheses, even if we owned the most incredible next-generation sequencing platform.

In traditional molecular biology things are a little bit more complicated. In the case of the article of Slug and Sox9, Weinberg's group derived the hypothesis that TFs might play an important role in determining and maintaining cellular states in adult stem cells from *previous studies* showing that kind of evidence for other stem cell states. From this, they derived the hypothesis that a similar mechanism might work also for adult stem cells. This kind of hypothesis is derived from previous knowledge, but the real inference is pursued by a mechanism of analogy. The idea is that as in certain types of stem cells there is such and such a mechanism*, then* we hypothesize that this might work also for adult stem cell. When they hypothesized a similar mechanism in breast cancer stem cells, they transferred their findings into a close field in order to derive a novel hypothesis. Also in the episode of the discovery of the mechanism of the protein synthesis, analogy and derivation from previous knowledge are ubiquitous. Zamecnik started his research by reasoning on the biochemical knowledge accumulated about polypeptides trying to hypothesize how they can be possibly assembled. Watson and Crick started their inquiry by reasoning on what they found about the double helix structure of the DNA and by using the analogy of the coding problem. Therefore it seems that hypothesis generation in traditional molecular biology is richer than in data-driven approaches.

There are some differences too in the two approaches to *prior assessment or hypothesis development*. In data-driven biology prior assessment is a composition of the elimination step of eliminative inferences and a development of hypotheses by means of consultation of databases, which should be seen as a way to consult prior knowledge of a field. In GWASs while statistical analyses eliminate less probable hypotheses, the use of

databases linked to big science projects such as ENCODE are a way to develop very abstract hypotheses such as "The SNP *x* has a role in the phenotype *y*" to "The SNP *x* has a role in the phenotype *y* by deregulating the gene *z*". The same case can be made for cancer genomics by replacing 'SNP' with 'somatic mutation'. In traditional molecular biology things get more complicated, but they do not stand differently. If you consider the episode of the discovery of the mechanism of protein synthesis, some hypotheses are developed and others are eliminated by adding new evidence, though the eliminative process is not as systematic as in data-driven projects. In the case of the article on Sox9 and Slug, once Weinberg and colleagues have derived a set of hypotheses by previous studies of the form "TFs *x*, *y* and *z* regulates mammary stem cell state", they used experiments to gain evidence of the fact that some of these TFs are not implicated in the phenomenon of interest, as well as putting details on the role of one of the TFs. But they used retroduction as well. In particular, when they do not know how to interpret concordant and discordant evidence about the role of Slug, they reasoned that one way of making sense of their data is to modify slightly their initial hypothesis. This is exactly a kind of abductive inference. In a sense, retroduction is present in data-driven biology as well. I mean that we make sense of statistical analyses by claiming that certain somatic mutations and SNPs exceed a particular statistical threshold because they are implicated in the phenomenon of interest in such and such a way. However, this step is less 'creative' than in traditional approaches. One of the criticisms to retroduction was that hypotheses (the premises) seem to come out of nowhere, while in data-driven biology there is a sort of 'assistant' (the database) guiding the refinement of hypotheses. Again, the regimented nature of data-driven biology makes this approach far more systematic than the traditional one.

Finally, *justification* takes place quite differently in the two approaches. In data-driven biology justification is a sort of low-level confirmation called 'validation', namely that practitioners provide some stringent experimental evidence that the entities identified in the previous phases actually play a role in a certain phenotype. In traditional approaches, as I have shown in Chapter 3, strong hypothesis testing is identified only retrospectively, in the sense that the fact that no further anomalies emerge from experiments (and hence evidence that can undermine productive continuity) is taken as a test that the mechanism description is sufficiently detailed and complete. Therefore, *justificatory phases in the two approaches are quite different*.

Let us now move the comparison to the level of epistemic values (Table 2).

	Hypothesis generation	**Prior assesment and development**	**Justification**
Traditional Molecular Biology	External consistency Internal consistency	External consistency Internal consistency Empirical adequacy	Internal consistency Empirical adequacy
Data-driven Molecular Biology	External consistency Internal consistency Empirical adequacy	External consistency Internal consistency Empirical adequacy	Internal consistency Empirical adequacy

Table 5.2. A comparison of the epistemic values employed in data-driven and traditional molecular biology within each phase of discovery

It goes without saying, internal consistency is pervasive at all levels of both approaches.

In *hypothesis generation* the only difference between the two approaches lies in valuing empirical adequacy by data-driven molecular biologists. As I have shown, in data-driven molecular biology the initial set of hypotheses depends on the consistency with the previous knowledge about SNPs, somatic mutations and certain phenotypes, but the members of the universe are 'negotiated' via empirical adequacy, in the sense that one has actually to detect a certain mutation by means of next-generation sequencing or SNP array to say that that mutation is a candidate for being implicated in the phenotype of interest. On the other hand, in 'traditional' molecular biology, one can generate a hypothesis merely by reasoning on previous studies, without having to connect hypotheses to evidence produced in her own lab. However, one can say that in a sense there is empirical adequacy too, since previous studies draw their conclusions from observations.

Next, the two approaches in *weak evaluation* and *justification* endorse the same kinds of values, namely external and internal consistency and empirical adequacy. As I

have shown for both data-driven and traditional molecular biology, elimination and development of hypotheses rest on both the contrast of hypotheses to actual observation, as well as on the comparison with previous studies or knowledge to refine biological claims. The phase of justification is based on how hypotheses fit actual observations and therefore it is guided by the desideratum of empirical adequacy.

2.1 Where is exactly the disagreement?

It seems that there are only four differences between the structures of discovery and in how the two approaches attack epistemic complexity. The first two are related to inferences, namely that 'traditional' molecular biology uses analogies to generate hypotheses and retroduction in prior assessment while data-driven biologists do not. The third is the endorsement of empirical adequacy by data-driven molecular biology in hypothesis generation. Finally, there is a difference in the way practitioners are aware of being in a justificatory phase. Should we say that the controversy is simply motivated on the basis of these differences? Actually, we might say that data-driven molecular biology employs an assembly-line version of the traditional approach. If we come back for a moment to decomposition and localization, everything becomes clear.

The first two phases of discovery of data-driven molecular biology can actually be subsumed into the 'mechanistic' perspective, in the sense that data-driven biologists attack epistemic complexity in a way that is completely compatible with the ones depicted by the mechanistic philosophy. Consider again the crucial distinction made by Bechtel and Richardson (2010) between localization and decomposition. These two strategies are considered as a starting point in mechanistic discovery. Decomposition "assumes that one activity of a whole system is the product of a set of subordinated functions" (2010 p 23) while localization tries to identify the entities that may play the subordinated functions. Clearly, there can be interplay between the two strategies especially in data-driven molecular biology. We may say that the particular problem *x* and the kind of solutions implied by the background assumption *y1, y2,…, yn*, lead to the assumption that the system investigated is somehow decomposable into subordinated functions. Accordingly, the system at hand is divided into several subcomponents *z1, z2,…, zn* (e.g. SNPs in a GWAS, somatic mutations in cancer genomics) that are claimed to be responsible for the subordinated functions, whatever these are. After the eliminative steps of eliminative inferences, some of the *zs* are retained as strongly associated with the phenomenon under investigation.

The interplay between decomposition and localization as I have just shown for data-driven molecular biology might be conceived in parallel to some remarks made by

Darden and Craver (2013). In discovering mechanisms, one looks immediately for entities or activities that might be involved in the phenomenon of interest. But the quest for mechanisms starts with a preliminary characterization of the phenomenon (precipitating conditions, modulating conditions, etc). From this preliminary idea about a phenomenon, one decomposes a system into parts, and then she identifies some as relevant. This is exactly the same in data-driven molecular biology. Again, given the preliminary characterization of a phenomenon *x*, and the kind of solutions implied by the background assumption *y1, y2,…, yn*, (where *x* and *ys* form the characterization of the phenomenon under scrutiny) the system at hand is divided into several subcomponents *z1, z2,…, zn* (e.g. SNPs in a GWAS, somatic mutations in cancer genomics) and some *zs* are retained as being relevant parts of the mechanisms producing the phenomenon.

Therefore, it seems that the discovery strategy of data-driven molecular biology is somehow compatible with the epistemic perspective provided by mechanistic philosophy. I am also tempted to say that data-driven molecular biology provides merely a particularly interesting version of the discovery strategies of molecular biology. The interesting part is that the data-driven approach provides a set of quite mechanical procedures to go through decomposition and localization. Data-driven molecular biology seems to be a sort of assembly-line instantiation of molecular biology, and it is just more efficient than traditional approaches. It is exactly what Dulbecco (1986) said when he was conjecturing about the cancer genome: Having a 'map' of the genes mutated can foster the discovery of cancer genes in a more efficient way than trial-and-error or piecemeal approaches. However, while data-driven biology promotes efficiency, its approach does not deviate from the general guidelines provided in the traditional *loci* of the literature on the discovery of mechanisms. This last remark has an important consequence. If both cancer genomics and GWASs are compatible with the discovery of mechanisms as described by 'mechanistic philosophers', and if the 'mechanistic philosophy' illustrates the research strategies employed also in traditional molecular biology, *then cancer genomics and GWASs (supposed to exemplify a new methodology for molecular biology) are neither in opposition to 'traditional' molecular biology, nor radically new*. Therefore, the epistemic perspective provided by 'mechanistic philosophy' can still make sense of many of the so-called data-driven biological studies. It seems that we should conclude that the controversy motivating the present work is completely misguided and misleading.

However, the attenuation of the conflict is just *illusory*. If we consider just inferences and traditional epistemic values within each phase of discovery the conflict is weakened; but if we look at the complete models, findings, hypotheses or research outputs of the two approaches, there are some remarkable differences. Weinberg-like

model[38] are particularly detailed and the decomposition and localization strategies seem to be extended *ad libitum*, providing mechanistic details until the slightest anomaly is removed. On the contrary, data-driven models seem to be quite 'skinny', with few mechanistic details but with an impressive statistical power. When in Chapter 4 I was discussing Bechtel and Richardson's notion of *direct localization*, I have emphasized that this task lies at a lower level than the identification of the locus of control because it segments the system on the inside rather than from the environment. However, Bechtel and Richardson also note that direct localization cannot really provide a detailed mechanistic description of why certain entities are involved in the development of a biological phenomenon. As a matter of fact, it is an insufficiently constrained localization, and it provides just correlation. It is very difficult to understand when localization is sufficiently constrained, but the take-home message of Bechtel and Richardson is that *we can always go downward with localization and decomposition* and the more we go downward, the better it will be for the mechanistic model. Data-driven biology seems to elaborate models that are *insufficiently constrained*. But data-driven models are insufficiently constrained *from the traditional molecular point of view*, and they are sufficiently constrained from another point of view. The *desideratum* of 'sufficiently constrained model' lies elsewhere; in particular it lies in the fact that their models must be projectable to more samples as possible. For a data-driven biologist the fact that a SNP seems to be implicated in the development of diabetes in 6,000 patients is more significant than finding the exact mechanism of action of that SNP in just a couple of cell lines. Therefore, we are back to epistemic desiderata: Data-driven and traditional molecular biologists seem to value different fundamental desiderata for their models. In order to pinpoint exactly this intuition, we should come back to the topic of epistemic values and try to understand precisely which kinds of values are at play besides traditional epistemic values such as internal/external consistency, empirical adequacy or simplicity.

3 EPISTEMIC, COGNITIVE AND QUASI-EPISTEMIC VALUES AND THEIR ROLE IN THE CONTROVERSY

In the first chapter I introduced the topic of epistemic values. I provisionally defined 'epistemic values' as epistemic desiderata and listed some traditional criteria for theory choice working as values. In a scientific context, an epistemic desideratum is said to be 'epistemic' because it is likely to promote those characters of science that make science

[38] For the sake of simplicity, here I use "models," "hypotheses" and "theories" as synonymously. A more precise approach would be to say that hypotheses are usually statements and theories are families of models. However I take a hypothesis, a theory or a model in the very basic sense of being a *x* which refers to a portion of the world, where this *x* is aimed at representing the portion of the world of interest in some sense to be specified.

the type of knowledge usually seen as "the most secure knowledge available to us of the world we seek to understand" (McMullin 1983, p. 18), while it is a 'desideratum' because it is something one believes will help to achieve that kind of knowledge, if adequately pursued. The core of McMullin's article (1983) was to show that theory appraisal is a procedure much closer to value judgement than to some rule-governed type of inference, and by 'value judgement' he means exactly *"the criteria of [theory] choice function not as rules, which determine choice, but as values which influence it"* (1977, p. 331).

However, the topic of science and values is – as one may expect – much more complicated. There is for instance a lively debate on the relation between scientific inferences and non-epistemic values (such as ethical and political values) stemming from an influential article by Rudner (1953) where he shows how non-epistemic values are embedded, inevitably, even in the *internal* procedures of science. This article also anticipates the contemporary debate on the role of non-epistemic values in scientific reasoning (Douglas 2000; Hempel 1965; Rooney 1992) as well as some strands of the philosophy of risk analysis. However, here I want to focus just on the so-called epistemic values.

3.1. Re-introducing epistemic values

The idea of 'values' in an epistemic sense as playing a role within the procedures of science is introduced explicitly by Kuhn (1977). However he does not mean values in an ethical sense. Kuhn's move is provocative. The fact that there are desiderata in theory choice is uncontroversial. Kuhn's move is to argue that criteria of theory choice work as values instead of being embedded in an unambiguous procedure governed by strict rules. His starting point on this issue is the final part of his *The Structure of Scientific Revolutions* (1962), where he writes that decisions of theory choice are not like proofs. The topic of theory choice is usually discussed by considering the typical and standard characteristics of a good scientific theory. Kuhn names five criteria: Accuracy (empirical adequacy), internal and external consistency, scope, simplicity, and fruitfulness. In Chapter 2 and 4 I have mostly taken into consideration some of these traditional criteria, without inquiring whether there can be other. If theory choice were a matter of proof and algorithms, then one would be able to apply these criteria without difficulties. However, in the application of these criteria - Kuhn continues - *there are* several problems, and a clear algorithm of theory choice is lacking. First, criteria are imprecise, in the sense that individuals usually do not agree on how these should be applied to concrete cases. Next, these criteria are repeatedly in conflict with one another, e.g. accuracy or scope (Kuhn 1977, p 357). Kuhn analyzes some examples in the history of science where conflicts

between uses of these criteria undermine the linearity of procedures of theory choice. Copernicus' system was far from being more accurate than Ptolemy's until Kepler drastically revised it, 40 years after Copernicus's death. Therefore, Kuhn says, there must be other reasons on the table to provisionally save and choose the heliocentric astronomy instead of Ptolemy's system. Another example is the phlogiston theory that matched experience in some areas better than oxygen theory, which in turn worked better in other areas of experience. Here the problem is: How do we decide which area of experience is more important for a criterion such as empirical adequacy? Kuhn's main point is that "the criteria of theory choice with which I began function not as rules, which determine choice, but as values which influence it" (1977, p 362). In other words the five criteria mentioned by Kuhn cannot dictate theory choice, but they are values (in the sense of maxims or norms) having a remarkable (but not conclusive) effect on theory choice. McMullin (1983) then refers to these values as 'epistemic' because they are supposed to be truth conducive. 'Truth-like' character is a sort of regulative ideal determining what should be counted as an epistemic value and what should be not (Rooney 1992).

Another seminal conceptualization of the issue of epistemic values stems from Longino's distinction between constitutive and contextual values roughly corresponding to the epistemic/non-epistemic distinction (1990; 1996). Constitutive values are taken to be those desiderata of good explanations, data, procedures, as well as hypotheses or theories. They are considered as desirable properties within scientific communities and they are generated from an understanding of the goals and aims of science. But, ideally, constitutive values have to be independent from 'contexts' and should be valued (or at least recognized as constitutive values) by all scientific communities. Another way to describe constitutive values is that desiderata such as accuracy, internal/external consistency and the like could be thought as "explicating what 'best' means in inference to the best explanation" (Longino 1996, p 44). Instead, contextual values are those values embedded in the social and cultural environment where science is actually done. These should be understood as political, ethical and cultural values.

3.2 Are epistemic values in data-driven and traditional biology truth-conducive?

Stemming from these traditional positions, some problems related to epistemic values *per se* arise. First, it is not clear what McMullin means by 'truth' or Longino by 'the aims of science'. If truth or the aims of science are taken to be the cornerstones for

discriminating epistemic and non-epistemic values and in general to identify epistemic values, then something more should be said about truth or aims. Second, some are suspicious about the epistemic nature of epistemic values. In the literature, epistemic or constitutive values are sometimes said to be 'cognitive values' (Laudan 1984; 2004). In particular, Laudan (2004) argues that epistemic values in science such as scope, generality and the like are not really *epistemic*. Laudan disputes that features associated to acceptable theories such as explaining known facts in a domain, or explaining different facts and so on are really associated to epistemology *tout court* in the sense of being truth-conducive *per se*. In particular, when speaking about 'rules of thumb' such as 'saving the phenomena' or 'consilience of inductions[39]' (which are clearly ubiquitous in theory appraisal), Laudan adds that it is neither necessary nor sufficient for the truth of a theory or a hypothesis that a theory or hypothesis maximizes any of these attributes. The fact that a theory cannot explain a fact is not an argument against its veracity. Therefore these virtues, Laudan continues (2004, p 18), are not really epistemic. Instead they deal with the breadth and the range of theories rather than their truth. Laudan calls these virtues/values 'cognitive values' because these criteria refer exactly to scientists' expectation about good theories that are not related to worries about their veracity. As Van Fraassen (1980), Laudan thinks that a theory does not have to be true to be good. Nonetheless, cognitive values "are constitutive of science in the sense that we cannot conceive of a functioning science without them, even though they fail to be intelligible in terms of the classical theory of knowledge" (Laudan 2004, p 19). Laudan makes an interesting argument by considering Kitcher's 'explanatory unification' (1993). Every unifying theory *T*, Laudan says, must clearly entail non-unifying counterparts *T1, T2,…, Tn*. All these non-unifying counterparts must be true if *T* is true. If *T* and its counterparts are all true but scientists regards *T* as better than its counterparts this is because for a virtue (the virtue of explanatory unification) that is non-epistemic, since counterparts do not possess it but they are not false or less true (whatever this means). Ergo, explanatory unification is not really epistemic.

In my analysis of epistemic values in data-driven and traditional molecular biology, I mentioned internal consistency, external consistency, and empirical adequacy. Let us try now to apply Laudan's argument on these three values. The theoretical 'virtue' of *internal consistency* is not really epistemic, in the sense that is not truth-conducive *per se*. However, we should admit that if a theory is not internally consistent, it is non-sensical. While internal consistency does not make a theory true, its absence indicates a serious problem for the theory at hand. In the empirical sciences, a similar analysis may be drawn for *empirical adequacy*. The absence of empirical adequacy points to serious problems of the theory in the first instance. Empirical adequacy is not truth-conducive *per*

[39] 'Consilience of induction' is the convergence of evidence toward a strong conclusion

se, but its absence indicates that a theory has serious problems. Therefore empirical adequacy and internal consistency are not really epistemic, but they should be conceived as indispensable virtues that a theory should embed in order to, at least, *not to be considered false before or after any process of theory appraisal*. Of course, a theory that is internally consistent and empirically adequate could turn out to be false anyway (as most theories are), but we cannot know this *a priori*. However, if a theory is not internally consistent and empirically adequate, such theory is not worth a second look.

Things stand differently for *external consistency*. If a theory is not externally consistent, this is not a hint of its falseness. A theory can be true even if it is not consistent with a bigger theory in its own field, especially because the bigger theory can be false. In general, external consistency does neither point to falseness nor to veracity. This criterion is more pragmatic in nature. For instance, if I elaborate a theory that seems to be true but is not consistent with the rest of the knowledge of the field, then things will get hard in defending my theory against criticisms, because critics will interpret the lack of coherence with the knowledge of the field as an attack to the *whole field*. If I elaborate a theory that is consistent with the rest of the field I am working, then the theory will attract fewer criticisms. We have seen both in data-driven and in traditional molecular biology that external consistency aids to generate and develop hypotheses, by drawing analogies or simple derivation from previous knowledge. Therefore, external consistency is also strategic (as well as pragmatic) in the sense that it aids strategies of discovery.

3.3 A taxonomy of cognitive values

The argument so far is that what have been traditionally called 'epistemic values' should not be called 'epistemic' because they are not directly truth-conducive. Rather, they should be called 'cognitive values' because they are related to scientists' expectations of features that a good theory should embed independently of its veracity.

The 'cognitive values' identified in data-driven and 'traditional' molecular biology should be divided in two groups. First (Group 1), there are criteria (empirical adequacy and internal consistency) whose absence indicates a serious epistemic problem. Second (Group 2), there are pragmatic and strategic values (e.g. external consistency). It is no surprise that data-driven and traditional molecular biology endorse the same values of the first group, because both tend to elaborate models that meet the minimal requirements for a theory not to be non-sense. It is no surprise either that both data-driven and traditional molecular biology embed external consistency – though in different phases of discovery – because the consistency of models with the knowledge of a field is a distinctive trait of the discovery strategies of molecular biology. Values of the first

groups are minimal requirements, while external consistency is a value applied to theories and models alone. The lesson to be drawn here is that differences between data-driven and traditional molecular biologists do not point to any of values of Group 1 and 2. Both traditional and data-driven endorse the minimal requirement for an empirical theory to be evaluated, i.e. internal consistency and empirical adequacy. Moreover, they both value external consistency, for pragmatic and heuristic reasons.

Differences in models of data-driven and traditional molecular biology point to another group of values (Group 3). Such values deal with theories/models and the kind of evidence one looks for in order to elaborate what she takes to be a good model. For instance, a mechanistic model is a good model if it points to many molecular details. According to traditional molecular biologists, data-driven models are not good models because they are not focused on enough molecular details. In this sense, the virtue of 'model rich in molecular details' that mechanistic models must embed is a criterion establishing in what respects a model/theory have to be similar to a portion of the world, as well as establishing the portion of the world itself we should be interested in. In order to embed the virtue 'model rich in molecular details' into a model, one has to investigate certain portions of the world by looking for molecular interactions and processes. In other words, values of Group 3 influence the type of evidence a scientist will look for in elaborating models and how a model should be similar to a specific portion of the world. Therefore, cognitive values behind models of data-driven and traditional molecular biology are values of theories/models in relation to portions of the world.

Now we have three groups of cognitive values: Minimal requirements, strategic, and values of theories/models when in relation to the portion of the real world that models have to be similar with. Douglas (2014) also groups cognitive values in three distinct sets. Though her first set corresponds to my Group 1, the other two are quite different from Group 2 and Group 3. Douglas' second set of values is one of pragmatic and strategic values as mine of Group 2. However, unlike Douglas, I do not attribute any epistemic import to these values. Her third set contains values applied to theories in relation to evidence. However, my Group 3 encompasses much more than 'evidence'. As a matter of fact, values of Group 3 set the way a model should be similar to some portions of the world.

To sum up, from my analysis emerges the existence of three distinct groups of cognitive values. First (Group 1), there are values which are minimal criteria for an adequate science. Absence of values such as internal consistency or empirical adequacy indicates a clear epistemic problem. In a sense, these cognitive values are the only epistemic values, at least indirectly. Failing to embed one of these values is a sign that a theory is false in any possible sense. Without these values, a theory cannot enter into the following phases of theory appraisal. These values do not contrast a theory with evidence

but they actually make the case that, if a theory does not embed them, then that theory cannot be compared to any evidence because of its inconsistent nature. Group 2 is composed by values which are strategic or pragmatic in nature and *they are applied to theories alone*. Finally (Group 3), there are values being desiderata of models/theories when applied to models/theories in relation to the portion of the world they have to investigate. I call these virtues *quasi-epistemic values*, in the sense that they provide a particular kind of genuine epistemic contribution. Quasi-epistemic values are particular desiderata of a model/theory when in relation with a portion of reality, in the sense that they are the kind of criteria we look at in order to establish a relation of similarity of a model to a certain state of affairs (Weisberg 2013). The label 'quasi-epistemic' stresses the modesty of my proposal with respect to McMullin's one: Quasi-epistemic values are not interested in *T*ruth, but rather in particular truths, and they direct models and theories towards particular aspects of the real world. *My claim is that data-driven and traditional molecular biology, while being similar with respect to inferences or epistemic values of Group 1, they substantially diverge in the endorsement of quasi-epistemic values (Group 3)*. Let us see this in detail.

3.4 Generality and depth as fundamental quasi-epistemic values of data-driven and traditional molecular biology respectively

Quasi-epistemic values are virtues of a theory considered in relation to a portion of reality. These values behave differently from values I have identified for the three phases of discovery highlighted in Chapter 1. My claim here is that quasi-epistemic values are not just criteria we value in judging a final hypothesis or theories against evidence. In a sense, *quasi-epistemic values constitute a sort of background guide throughout the whole process of discovery*. If we value above all predictability, then we will generate, develop and strongly evaluate hypotheses or models that will contain more predictability features *independently from the actual inferences and epistemic values of Group 1 and 2 employed in hypothesis generation, hypothesis development and strong evaluation*.

Here my analysis departs substantially from Douglas' view. The aim of Douglas' article (2014) is to provide an epistemic foundation for why some cognitive values are pervasively listed. In other words, Douglas wants to pinpoint epistemic reasons why some cognitive values are so important. This concern stems from the idea that "some set of values is (by and large) what has been important to scientists in their practice, and that should be good enough for philosophers of science" (Douglas 2014). While this line of research is valuable, my idea is that such a foundational analysis can be a limiting factor in identifying cognitive values themselves, especially quasi-epistemic values. The fact is that theoretical virtues should not be restricted to the traditional lists made by McMullin

or Kuhn. Longino (1995) tries to contrast the typical list of cognitive values with an alternative list drawn from feminist studies of science. Therefore here I do not want to limit my analysis of the identification of quasi-epistemic values to the list made by Douglas[40]. But how do I identify quasi-epistemic values that set the aims and the scope of data-driven and traditional molecular biology? This is straightforward: I will look at those desirable properties that a complete theory or hypothesis should have within the communities of data-driven and traditional molecular biologists.

Let me start from *traditional molecular biology*. As I emphasized, the aim of molecular biology is to elaborate detailed descriptions of mechanisms. Mechanisms are entities and activities organized in such a way that their interactions produce a biological phenomenon. Molecular biologists are thus interested in the molecular details to fill black boxes in a mechanism sketch. We have seen in the episode of the discovery of the mechanism of protein synthesis that scientists were mainly interested in filling the gaps in the productive continuity of their models. Also in Weinberg's article, it is clear that his group tries to make sense of a certain phenomenon by identifying the details of the molecular interactions of several entities, identified through multiple experiments. If we look at this idea of 'identifying molecular details' in light of the interplay between decomposition and localization the aim of molecular biologists turns out to be clear. Each round of decomposition and localization is exactly an attempt to go lower in the level of analysis. Anytime a molecular biologist designs an experiment is actually using the strategy of decomposition and localization to go *deep* into the molecular details. The more molecular details a mechanistic model has, the better it is. But if we need to go lower in the level of analysis to get molecular details, and if we need a progressive downward decomposition/localization, then the criterion to develop models in traditional molecular biology is what I would like to call *depth.* The desire and the will to go *in depth* with decomposition and localization is the aim of traditional molecular biology, and we value a model as sufficiently adequate if it goes sufficiently *in depth in molecular details*. By valuing such models, we think that the relation connecting a truth-bearer (the mechanistic model) to the portion of reality is one focusing on molecular details, and the deeper we go, the better our model approximates to truth. To summarize, the quasi-epistemic values *par excellence* of traditional molecular biology is *depth*. This makes sense also if we consider *productive continuity* (Darden 2006): A good mechanistic description (and hence a good explanation) must show exactly how *each stage* of the mechanistic chain produces the next one, i.e. improving the quality of an explanation means filling gaps in a chain connecting set-up and terminations side of the causal chain. This means that we should look for more molecular details to reach productive continuity,

[40] To tell the truth, Douglas explicitly says that her list is not meant to be exhaustive

and hence we should go lower in the level of analysis with the interplay between decomposition and localization.

Let us now move to data-driven biology. As I have already stressed, from a traditional molecular biology point of view data-driven biologists seem to elaborate insufficiently constrained models. It seems that data-driven biology level of granularity stops at the level of *direct localization*, or just after that. Therefore from Weinberg's point of view, data-driven biologists elaborate models which simply fail to be similar enough to the portion of reality that molecular biology is interested in. From this perspective, such models are simply sloppy. However, data-driven biologists are clearly interested in something else. If you consider GWASs, there is the perception that practitioners are interested in the statistical power of their results, so that their findings can possibly be applied to as many patients in the real world as possible. A similar situation applies to cancer genomics. The aim of consortia like TCGA is to do screenings with an increasing statistical power so that their findings (e.g. the identification of driver mutations) can potentially apply to as many cancer genomes as possible. If the aim of data-driven biology is to apply their models to more objects in the world, we may say that they value *generality*. The concept of 'generality' in the literature on scientific models is pervasive. I find particularly useful the characterization of the concept of 'generality' made by Michael Weisberg, especially in (Weisberg 2007; Matthewson and Weisberg 2009). In a preliminary sense, generality can be understood as "a measure of how many phenomena a model or set of models successfully relate to" (Matthewson and Weisberg 2009, p 180). Of course, one should need to specify exactly what is the model-world relation. One may mention isomorphism (van Fraassen 1980), partial isomorphism (da Costa and French 2003), or other notions of similarity. Of course I cannot enter in this debate. I just want to argue that data-driven biologists are interested in elaborating models that relate to as many phenomena as possible, independently of the notion of similarity model-world they have in mind. However, Weisberg distinguishes between two types of generality, and this distinction turns out to be very useful for the purpose of this chapter. Accordingly, generality can be understood as the measure of how many *actual* targets our models relate to (*a-generality*) or as a measure of how many *possible* targets our models could in principle relate to (*p-generality*). Studies such as GWASs or cancer genomics screenings elaborate models that have a high measure of a-generality. But the important point here is the statistical power. The idea of data-driven biologists is that, the more samples we have, and the more statistically robust our findings (targeted mutations, genes, etc) are, the more promising a mutation or a gene can be as a target for drug discovery studies. In other words, the more a-general a model is, the more p-general it might be. Therefore data-driven biologists are focused on increasingly big numbers because they want to achieve a high measure of a-generality as they think that, the more a-general, the more

p-general, and the more p-general means that they are unveiling some robust and crucial components of complex phenomena such as tumors or chronic diseases. This is the attitude guiding and shaping their context of discovery.

Now the controversy turns clear. Traditional and data-driven molecular biologists have similar discovery structures and inferences because data-driven biology is a more regimented version of the traditional strategies. They also endorse similar epistemic values of Group 1 because these values should be endorsed by *any* scientific theory/hypothesis/model. They value also external consistency (Group 2) because this pragmatic desideratum is a hallmark of the molecular biological tradition. Therefore, focusing the analysis just on this group of values is uninteresting. However, the real disagreement lies in the endorsement of different quasi-epistemic values, which shapes the whole process of discovery and also explains why these communities elaborate different models.

4. CONCLUSION

In this chapter, I compared the discovery strategies of traditional and data-driven molecular biology. Drawing from my analyses on Chapter 2 and 4 based on the tripartite framework elaborated in Chapter 1, I have compared inferences and (preliminary called) epistemic values within each phase of discovery of the two research traditions. However, a preliminary result was that differences in inferences and traditional epistemic values could hardly motivate the heated disagreement between traditional and data-driven molecular biologists. Moreover, it seemed that data-driven biology was merely a sort of assembly line version of traditional molecular biology.

Next, in analyzing models and final hypotheses of the two sides, I pointed out that data-driven and traditional biologists seem to value different desiderata or representational ideals. Therefore I had a closer look at the issue of epistemic values. In scrutinizing traditional and more recent loci of this literature (Kuhn 1974; McMullin 1983; Longino 1990; 1995; 1996; Laudan 1984; 2004; Douglas 2014) I have emphasized (a) that "epistemic" values are not really epistemic, (b) that there is a broader group of values called 'cognitive values' and the group of epistemic values is just a relatively uninteresting subgroup of it, (c) that cognitive values can be divided in epistemic, pragmatic/strategic and a third group and (d) that the third group is composed by values understood as desiderata of theories evaluated against evidence, which I have called *quasi-epistemic values* in order to highlight the modesty of my proposal with respect to

McMullin's one. I have also emphasized that the cognitive values I have identified in Chapter 2 and 4 are mostly epistemic, and since these values are endorsed necessarily by any scientific theory/hypothesis/model then they are not so relevant to identify differences between research strategies. I then argued that the disagreement of the two sides lies in the endorsement of different quasi-epistemic values. I have made the cases that traditional molecular biologists tend to value *depth*, while data-driven practitioners are more focused on *a-generality* because they think that a-generality can be a shortcut for *p-generality*.

To conclude, now it is possible to understand the nature of the disagreement. When Weinberg says that data-driven approaches cannot really achieve the typical results of molecular biology he is arguably right. Since data-driven biologists are interested in generality, their interplay of decomposition and localization will stop at a coarser-grained level of analysis, i.e. at a level of direct localization. While in the mechanistic tradition direct localization can just provide correlation and it is taken to be insufficient for elaborating *mechanistic details*, in the developing data-driven tradition mechanistic details are not so important, in the sense that once we have reached a sufficiently high a-generality, then we are entitled to infer p-generality which is the most important criterion for theory choice in this context. On the other hand data-driven biologists would appraise negatively an article such as the one of Weinberg's group analyzed in Chapter 4, because their results apply to few cell lines and few patients. In other words, Weinberg's group has elaborated a model that is insufficiently constrained from the point of view of data-driven biology because it scores low in a-generality and, as a consequence, in p-generality. Therefore the disagreement is motivated by the fact that each community applies its prominent desideratum to the results of the opposed community, which are shaped by a different quasi-epistemic value. In other words, there is a problem of communication because traditional and data-driven biologists use the same words (e.g. reliable results, reliable models) but with different meanings (e.g. 'a reliable model is one that goes sufficiently deep in the mechanistic details' versus 'a reliable model is one that has a high measure of a-generality').

CHAPTER 6

CONCLUSIVE REMARKS AND OPEN QUESTIONS

1. Summary of the main arguments

The disagreement between traditional and molecular biologists lies exactly in the type of insights that both 'traditions' actually achieve, and which types of insights are the best for biology or for the progress of biology. Data-driven and traditional molecular biologists endorse different criteria of theory choice or, as I called them in Chapter 5, different *quasi-epistemic values*. This type of disagreement is, at first sight, impossible to be disentangled. As in the controversy opposing Bohr and Einstein, here two different communities disagree on the very notion of 'virtue' of a scientific theory.

One might say that this is just an armchair discussion. After all, the most important thing in science is, trivially, to do science. This type of consideration discloses a quite naïve conception of science, something detached from the complexity of research agendas. The truth is that the controversy motivating the present work is not just a technical quarrel between rich academics debating from their 'ivory towers'. There is something more on the table, something that can make the difference in the 'main street'. Especially in the field of molecular oncology - where the (translational) link between basic and applied research is essential - the controversy is about which kinds of insights can better serve the aim of biomedicine. And since the label 'biomedicine' is a bottomless pit where a significant part of molecular biologists take money to do basic research, then the controversy between data-driven and traditional molecular biologists becomes significant because, instead of asking "What is the best methodological approach to molecular biology?", one might ask "What is the best methodological approach (serving the aims of biomedicine) that we should fund?". This is exactly what Bruce Alberts (2012), implicitly, asked in his *The End of Small Science*. Therefore there are several open questions coming out of my analysis. Before going deeper into this, let me sum up the main clams of my dissertation chapter by chapter.

The title of this dissertation is *The Context of Discovery of Data-driven Biology*. Now it should be clear that by 'context of discovery' I mean the sum of the three phases constituting my framework of discovery, the types of inferences used within each phase, the cognitive values endorsed, and in particular the quasi-epistemic values guiding the whole process. Since my aim was to compare the discovery strategies of two different research traditions (or, at least, two different research traditions according to a specific received view) in Chapter 1 I have analyzed the debate on discovery in science in the

philosophy of science starting from the distinction between the context of discovery and the context of justification. Clearly my notion of 'context of discovery' has nothing to do with the notion of 'context of discovery' put forth by philosophers such as Reichenbach. In the first chapter I made an historical and critical reconstruction of the debate on scientific discovery in philosophy of science arguing that discovery in science is a process composed of three phases: Hypothesis generation, hypothesis development and strong evaluation. My argument relies on the fact that these three phases are the minimum set of phases that a scientist goes through when discovering. Of course, there might be other sub-phases as well. However, for the sake of my argument I wanted just to identify the main epistemic moves in discovery. From this, my plan was to identify these three epistemic phases in both traditional and data-driven molecular biology in order to compare the two approaches on a common ground.

In Chapter 2 I have analyzed the structure of discovery of data-driven biology. I have grounded my analysis on two classes of case studies, namely genome-wide association studies and cancer genomics screenings. My findings on this matter is that data-driven molecular biology has an eliminative inferential structure, complemented by a comparison of data with data stored in biological databases in order to develop hypotheses not eliminated by statistical evidence. After the phase of hypothesis generation (that is strictly connected to background assumptions theoretically informing practitioners), an initial universe of abstract hypothesis of the form "the SNP/mutation/gene *x* has a causal role in the phenotype *y*" is narrowed by means of statistical analysis. After that, the universe of remaining hypotheses is then scrutinized by means of a computational analysis in order to develop hypotheses of the form "the SNP/mutation/gene *x* has a causal role *z* in the phenotype *y*". Clearly, the 'z' is very important, since if 'z' has nothing to do with the phenotype of interest (I have made the example of olfactory receptor genes and lung cancer), then the hypothesis is discarded. Finally, in the validation phase, experimental evidence is provided in order to strongly evaluate well-developed hypotheses. Although data-driven molecular biologists do not elaborate detailed mechanistic descriptions, they nonetheless move within the epistemic horizon depicted by mechanistic philosophers. In particular, data-driven molecular biologists provide strong statistical evidence for direct localizations of various sorts. Next I have identified the cognitive values guiding each phase of data-driven molecular biology.

Chapter 3 should be conceived as complementary to Chapter 2. Since in Chapter 2 I have discussed sparsely the epistemic importance of biological databases in data-driven discovery strategies, in Chapter 3 I have analyzed the epistemic status of databases in detail. In one sense, biological databases are idealized tools used to identify relevant phenomena in a data set by means of comparison: Biological databases are *evidence*

enhancers. In another sense, biological databases are explored with theoretical aims in mind, namely to compile highly constrained theoretical stores of the biological field.

In Chapter 4, I turned to 'traditional' approaches in molecular biology. First, I tried to delineate aims and scope of what is called 'molecular biology'. I have then limited my analysis of molecular biological discoveries to the mechanistic strategies used to attack the epistemic complexity of the molecular realm. I reconstructed Bechtel and Richardson's epistemological analysis of mechanistic discovery (2010) as well as Lindley Darden's (2006; Craver and Darden 2013) indispensable mechanistic philosophy (especially the epistemological side). I have then illustrated these strategies by discussing the history of the discovery of the mechanism of protein synthesis. Since in Chapter 2 I discussed cases related to cancer studies, and since one of the main characters of the controversy motivating this work is Robert Weinberg, in Chapter 4 I turned my attention to molecular oncology. I have discussed the origin of the molecular biological approach in oncology (Weinberg 2014) and I have analyzed an article coming out of Weinberg's lab (Guo et al 2012). I have shown how this study fits very well both Bechtel and Richardson/Darden's approaches, as well as my tripartite framework. Finally, I identified the main cognitive values driving such discovery strategies.

In Chapter 5 I compared 'traditional' and data-driven molecular biology. However, in the first instance the analysis has turned out to be disappointing. The preliminary result has been that the few differences in inferences and epistemic values could hardly motivate the controversy. However, at a finer-grain level of analysis something turned out to be of some interest. While the data-driven approach is a sort of assembly-line version of the discovery strategies depicted by so-called mechanistic philosophers, it seems that data-driven biologists restrict the interplay between decomposition and localization to the level of direct localization. Therefore, here is a hint of the disagreement. According to a traditional molecular biologist, results provided by data-driven molecular biologists are *insufficiently constrained localization*, and this judgement can take the form of 'correlation is not enough to establish the details of a mechanisms or causal production'. However, from the point of view of a data-driven biologist, it is exactly the kind of result provided by traditional biologists that are insufficiently constrained. One can elaborate the most detailed mechanistic description for a biological phenomenon, but if that description is derived only from the analysis of a couple of cell lines or a few samples, then the statistical power of this description is insufficient. Therefore it seemed that there are different criteria of theory choice at hand here. For these reasons, I analyzed more in detail the topic of epistemic values than I did in Chapter 1 and I have elaborated – by means of some modifications of the received view in literature, in particular (Douglas 2014) – a taxonomy of criteria of theory choice that have been called in the last few years *cognitive values*. Cognitive values are different if we identify properly

their targets. They can be applied to theory alone, to theory in relation to empirical evidence and to minimal requirements that any theory should meet. Epistemic values are just a proper subset of cognitive values. Moreover, it is not a very interesting set. As I have shown, data-driven and traditional molecular biology endorse the epistemic values of empirical adequacy and internal consistency, but just because these are the minimal requirements for a theory just to preliminary make sense. Therefore, looking for reasons of disagreement within this set of values is idle. Next, I identified the real disagreement of the controversy motivating this dissertation in the endorsement of different cognitive values of theories with respect to evidence. I called these values *quasi-epistemic values* to highlight the modesty of my proposal with respect to McMullin's one: These values are truth-conducive not in the sense that they lead to Truth, but just that they restrict the interest of investigators to certain aspects of truths or to some 'interesting' truths. Moreover, I have also added that these values constitute a sort of background guiding the whole process of discovery. Traditional molecular biologists tend to endorse *depth*, in the sense that they are interested in going lower in the interplay between decomposition and localization, in order to pinpoint as many mechanistic details as possible. On the other hand data-driven biologists are more focused on the generality of their results, defined as "a measure of how many phenomena a model or set of models successfully relate to" (Matthewson and Weisberg 2009, p 180). Drawing from the distinction between a-generality and p-generality delineated by Matthewson and Weisberg (2009), I concluded that data-driven biologists are interested in a-generality because they think that a-generality is a good indication of p-generality. To sum up, the disagreement lies exactly in endorsing and valuing two different criteria for theory choice, namely *depth* and *generality*. Therefore the disagreement is motivated by the fact that traditional molecular biologists interpret data-driven results in light of their quasi-epistemic value. In other words, if for traditional biologist a reliable model is a model that goes sufficiently deep in the mechanistic details, for data-driven biologists a reliable model is a model that has a high measure of a-generality. This is the core of the epistemological controversy.

2. THREE OPEN QUESTIONS

My results leave room for, *at least*, three open questions. Two are related to depth and generality, while one is strictly connected to the use of databases and to the assembly-line nature of data-driven biology.

About depth and generality, two questions are in order: (1) what is the relation between depth and generality? Do they stand in a tradeoff relation? (2) What is the best quasi-epistemic value for biomedicine?

On the other hand the assembly-line features of data-driven molecular biology poses normative questions over the constraints on discovery strategies and how such constraints are *attempts of unification of biological knowledge through bioinformatics tools*. Let us see each of these questions in detail.

2.1 Relation between depth and generality

In philosophy of science there have been some attempts to understand the relation between theoretical virtues – here understood in the same sense of cognitive values (Levins 1966; Weisberg 2004; Matthewson and Weisberg 2009; Odenbaugh 2003; Douglas 2014). The need to study such relations stems from the fact that scientists cannot build models embedding all the representational ideals that a model can possibly exemplify (Matthewson and Weisberg 2009). This can be a consequence of tradeoff relations between cognitive values. Standing in a tradeoff relation means that there is a relationship of attenuation, which occurs "when an increase in the magnitude of one attribute makes the achievement of another more difficult" (Matthewson and Weisberg 2009, p 170). Matthewson and Weisberg then analyze the kind of tradeoff relations that can possibly hold between precision and generality. They distinguish several types of tradeoffs. For instance there is strict tradeoff, namely the fact increasing the magnitude of one automatically decreases the magnitude of the other. The second is the 'increase tradeoff' occurring when it is impossible to increase or decrease simultaneously the magnitude of two representational ideals. Finally, there is 'Levins' tradeoff', namely that it is impossible to maximize both attributes of interest. Tradeoff relations are just an example. Douglas (2014) tries to uncover other types of relations, by dividing cognitive values in several groups. The ideas behind these types of works is to understand whether some cognitive values are mutually irreducible, and how certain scientific controversies involving the endorsement of different theoretical virtues between parties can be somehow disentangled.

The same could be done also for the controversy between traditional and data-driven molecular biologists. How can we understand the relation between depth and generality? Is there a tradeoff relation? If yes, which kind of tradeoff relation? Once we have uncovered the relation between those values, one could ask: Is that relation the result of pragmatic considerations or it is a consequence of the logic of representations at hand?

2.2 Depth, generality and biomedicine

Let us assume that there is a tradeoff relation between depth and generality, say strict tradeoff. How do we establish what is best desideratum for molecular biology? Clearly, it is very difficult to establish a sort of epistemological foundation explaining why depth and generality have become so important for molecular biology. Pure epistemic reasons are intertwined with pragmatic considerations and socio-historical analyses. Therefore, it is very difficult to understand which quasi-epistemic value should be endorsed in principle. But it is possible to delimit this task by contextualizing a little biological research. Here I restrict my analysis to the molecular biological works embedded in the biomedical agenda. Therefore, it is this agenda that we should look at to settle the controversy. This consideration can be linked with two sources of inspiration. First, let us consider an article by Helen Longino (1995) where she contrasts a typical list of theoretical virtues (accuracy, internal/external consistency, simplicity, fruitfulness, etc) with a feminist list of cognitive values (novelty, ontological heterogeneity, applicability to human needs, etc) in order to show that, sometimes, the only reasons for preferring a traditional instead of a feminist value (and vice versa), are eminently socio-political. Therefore it is not just that non-epistemic values can function as epistemic, but also that non-epistemic values can settle epistemic quarrels. The idea is that even "the apparently neutral criteria of accuracy or empirical adequacy can involve socio-political dimensions in the judgement of *which* data a theory or model must agree with" (Longino 1995, p 396). Therefore in order to understand which value between depth and generality is more appropriate we should look at the biomedical context, and the role that molecular biology has within the biomedical agenda. But the task here is not to explain how these values have emerged given the certain socio-political context of biomedicine and its goals, but rather to understand which value is the most appropriate to serve the duties that molecular biological basic research have in the biomedical context. Therefore, it is a *normative inquiry*, not (or not merely) descriptive. Clearly, I cannot develop on this further, but let us just say that a particularly obvious (but not trivial) duty of molecular biology within the context of biomedicine is to provide target molecules for drug discovery. Of course, there can be others, but let me simplify the context to make a point. These target molecules should be as much promising as possible in terms of being 'druggable'. Therefore the question is: Which quasi-epistemic value can promote an efficient molecular biology in the context of biomedicine (efficient in the sense of discovering more promising target molecules for drug discovery)? Is the approach valuing statistical powers of findings, or the approach having as its main goal the construction of detailed mechanism

descriptions? Are these mutually exclusive? Please pay attention to the fact that these questions are important because they set the basis for the answer to other questions: Which projects should we fund? How do we distribute money among small science labs and big consortia (e.g. hundreds of small labs or the 100,000 Genome Project)? Therefore what is needed is a clear organization of scientific inquiry having clearly in mind the biomedical agenda, which is for the most part funded by public money. In other words (and here we come to the second source), what we need here is something similar to the idea of *well-ordered science* (Kitcher 2001).

2.3 Tools for unifying biology

The last open question is related to the idea that data-driven molecular biology is a sort of assembly line version of traditional molecular biology. We have seen its division of labor between wet lab biologists, computational biologists, and biological maps (about maps, see Boem 2015). This has somehow implied *standardizations* not only of practices (e.g. strict division of labor, strict protocols, a selected number of platforms for sequencing etc) but also a standardization of biological knowledge. This can be thought in line with the notion of *regulatory objectivity* (Cambrosio et al 2006). This refers to the idea that objectivity in science is somehow generated as well as constrained by specific forms of collective production of evidence where collective refers to the evidence produced by, for instance, research consortia developing collective devices guiding the processes of discovery. In other words, what counts as 'objective' is not "whether or not the results produced by a particular laboratory are true, in some absolute sense, but whether or not they are compatible (within conventionally determined statistical limits) with results produced by other laboratories" (Cambrosio et al 2006, p 192). Putting aside socio-constructivist drifts, what is to be considered here is that there are certain practices by big consortia that systematically pigeonhole data into predetermined categories, as well as determining the way in which these categories should be conceived and combined. Consider what I have shown about biological databases in Chapter 3. The fact that we identify certain biological phenomena in data sets depends strictly on whether the shape of data is sufficiently similar to archetypal phenomena type stored in a sort of computational hyperuranion. But the way this 'cyber-hyperuranion' is conceived and built in turn depends on the practices of research consortia such as the Genome Reference Consortium or The ENCODE Project Consortium. Therefore evaluation of evidence is an operation distributed across the single scientist interpreting data and computational infrastructures. A similar argument can be made also for the Gene Ontology Consortium (2000). The grand project of Gene Ontology is motivated by the fact that biological

knowledge is sparse, and often biologists use the same words with different meanings (Boem 2015). Therefore Gene Ontology aims to provide a controlled vocabulary that can constitute a common framework for biologists (though in the dynamical context of the progress of biological knowledge). Gene Ontology is conceived as a tool for the unification of biology. If you consider the biological criteria to develop hypotheses in GWASs and cancer genomics shown in Chapter 2, the 'functional analysis' – that is, the identification of function – is based on Gene Ontology functional categories. I have also shown how this functional analysis is fundamental to develop hypotheses in a way or another. This is not to say that 'knowledge' is completely dependent on the social dimensions of the consortia creating the standard of knowledge. As a matter of fact, databases and Gene Ontology categories can be conceived as a sort of interface connecting scientists to data sets. It is a level of abstraction or, due to the complexity of these computational complexes, a *gradient of abstraction* (Floridi 2008). After all, this is very similar to the Darden's idea that the store of a field constrains the way discovery is done as well as what is discovered. But in Darden's case the store of a field is something sparse that one may find in textbooks, or can be incomplete due to lack of the knowledge of the single scientist. Here the situation is different. Scientists are necessarily connected to the whole corpus of biological knowledge, independently of their understanding of that biological knowledge. This kind of standardization transform what counts as an adequate explanation within the molecular biological field, as well as shaping how things are discovered. But this is not just integration, if we understood integration as putting together perspectives or methodologies in order to solve specific and local problems (Plutynski 2013). Rather, these tools seem really to unify molecular biology well beyond specific necessities typical of problem solving. Therefore it would be interesting to develop a notion of *unification* (decoupled from the notion of explanation (Morrison 2000)) shaped by the idea of standardization, and to track down how (and in which sense) standardization unify the field, the epistemic dynamics involved in this process of unification, and the consequences for scientific discovery and the progress of the whole field.

REFERENCES

Achinstein, P. (1970). Inference to scientific laws. In R. Stuewer (Ed.), *Historical and Philosophical Perspectives of Science* (pp. 87–104). University of Minnesota Press.

Achinstein, P. (1971). *Law and Explanation*. Oxford: Oxford University Press.

Achinstein, P. (1987). Scientific discovery and Maxwell's kinetic theory. *Philosophy of Science*, *54*.

Alberts, B. (2012). The End of " Small Science "? *Science*, *337*(September), 1230529.

An, O., Pendino, V., D'Antonio, M., Ratti, E., Gentilini, M., & Ciccarelli, F. D. (2014). NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database*, *2014*, bau015–bau015. doi:10.1093/database/bau015

Bechtel, W., & Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441. doi:10.1016/j.shpsc.2005.03.010

Bechtel, W., & Richardson, R. (2010). *Discovering Complexity - Decomposition and Localization as Strategies in Scientific Research*. Cambridge, Massachusetts, and London, England: The MIT Press.

Biankin, A. V, Waddell, N., Kassahn, K. S., Gingras, M.-C., Muthuswamy, L. B., Johns, A. L., … Grimmond, S. M. (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, *491*(7424), 399–405. doi:10.1038/nature11547

Boem, F. *A Matter of Style – How Map Thinking and Bio-ontologies Shape Molecular Research*. PhD Thesis. University' degli Studi di Milano

Bogen, J., & Woodward, J. (1988). Saving the Phenomena. *The Philosophical Review*, *XCVII*(3).

Bogen, J., & Woodward, J. (1992). Observations, theories and the evolutions of the human spirit. *Philosophy of Science*, *59*(4).

Bogen, J., & Woodward, J. (2005). Evading the IRS. In M. Jones & N. Cartwright (Eds.), *Poznan Studies in the Philosophy of Science and the Humanities* (Idealizati., pp. 233–268). Amsterdam: Rodopi.

Boniolo, G. (2013). On Molecular Mechanisms and Contexts of Physical Explanation. *Biological Theory*. doi:10.1007/s13752-012-0073-z

boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society*, 1*5*(5).

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. a, Kasowski, M., … Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790–7. doi:10.1101/gr.137323.112

Braithwaite, R. (1953). *Scientific Explanation*. Cambridge University Press.

Brookfield, J. F. Y. (2010). Q&A: promise and pitfalls of genome-wide association studies. *BMC Biology*, *8*, 41.

Buchanan, B. G. (1982). Mechanizing the Search for Explanatory Hypotheses. In *PSA : Proceedings of the Biennial Meeting of the Philosophy of Science Association , Vol . 1982, Volume Two : Symposia and Invited Papers (1982)*.

Burian, R. M. (1996). Underappreciated pathways toward molecular genetics as illustrated by Jean Brachet's cytochemical embryology. In S. Sarkar (Ed.), *The philosophy and history of molecular biology: New perspectives* (pp. 67–85). Dordrecht: Kluwer.

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

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

Cancer, T., & Atlas, G. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, *489*(7417), 519–25. doi:10.1038/nature11404

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

Cartwright, N. (1989). *Nature's Capacity and their measurement*. Oxford: Oxford University Press.

Church, D. M., Schneider, V. a, Steinberg, K. M., Schatz, M. C., Quinlan, A. R., Chin, C.-S., … Flicek, P. (2015). Extending reference assembly models. *Genome Biology*, *16*(1), 13. doi:10.1186/s13059-015-0587-3

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, *45*(10), 1127–1133. doi:10.1038/ng.2762

Cortes, A., & Brown, M. a. (2011). Promise and pitfalls of the Immunochip. *Arthritis Research & Therapy*, *13*(1), 101. doi:10.1186/ar3204

Craver, C., & Darden, L. (2013). *In search of Mechanisms*. Chicago: The University of Chicago Press.

Craver, C., & Darden, L. (2013). *In search of Mechanisms*. Chicago: The University of Chicago Press.

Curd, M. (1980). The logic of discovery: three approaches. In T. Nickles (Ed.), *Scientific Discovery: Logic and Rationality* (pp. 201–219). Dordrecht: Reidel Publishing Company.

Da Costa, N., & French, S. (2003). *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning*. Oxford: Oxford University Press.

Darden, L. (1982). Artificial Intelligence and Philosophy of Science : Reasoning by Analogy in Theory Construction. In *PSA : Proceedings of the Biennial Meeting of the Philosophy of Science Association, Vol . 1982, Volume Two : Symposia and Invited Papers*.

Darden, L. (1991). *Theory Change in Science: Strategies from Mendelian Genetics*. New York: Oxford University Press.

Darden, L. (2006). *Reasoning in Biological discoveries*. Cambridge, Uk: Cambridge University Press.

Darden, L. (2006). *Reasoning in Biological discoveries*. Cambridge, Uk: Cambridge University Press.

Darden, L. (2009). Discovering Mechanisms in Molecular Biology: Finding and Fixing Incompleteness and Incorrectness. In J. Meheus & T. Nickles (Eds.), *Models of Discovery and Creativity*. Dordrecht: Springer.

Darden, L., & Craver, C. (2002). Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *33*, 1–28.

Darden, L., & Maull, N. (1977). Interfield Theories. *Philosophy of Science*, *44*.

Darden, L., & Tabery, J. (2010). Molecular Biology. In *Stanford Encyclopedia of Philosophy*.

Davis, C. F., Ricketts, C. J., Wang, M., Yang, L., Cherniack, A. D., Shen, H., … Creighton, C. J. (2014). The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*, *26*(3), 319–30. doi:10.1016/j.ccr.2014.07.014

Dolnik, A., Engelmann, J. C., Scharfenberger-schmeer, M., Mauch, J., Kelkenberg-schade, S., Haldemann, B., … Bullinger, L. (2015). Commonly altered genomic regions in acute myeloid leukemia are enriched for somatic mutations involved in chromatin remodeling and splicing, *120*(18), 83–93. doi:10.1182/blood-2011-12-401471.

Douglas, H. (2014). The Value of Cognitive Values. *Philosophy of Science*, *80*(5).

Douglas, H. E. (2009). Reintroducing Prediction to Explanation. *Philosophy of Science*, *76*(4), 444–463.

Douglast, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, *67*(4), 559–579.

Duhem, P. (1955). *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press.

Dulbecco, R. (1986). A Turning Point in cancer research: Sequencing the Human Genome. *Science*.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, Massachusetts, and London, England: The MIT Press.

Eddy, S. R. (2013). The ENCODE project: Missteps overshadowing a success. *Current Biology*, *23*(7), R259–R261. doi:10.1016/j.cub.2013.03.023

Elliott-graves, A., & Weisberg, M. (2014). Idealization, *3*, 176–185.

Ellis, M. J., Ding, L., Shen, D., Luo, J., Suman, V. J., Wallis, J. W., … Mardis, E. R. (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*, *486*(7403), 353–60. doi:10.1038/nature11143

Fearon, E. R., & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, *61*.

Fernández-Suárez, X. M., Rigden, D. J., & Galperin, M. Y. (2014). The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic Acids Research*, *42*(1), D1–6. doi:10.1093/nar/gkt1282

Figueroa, M. E., Abdel-Wahab, O., Lu, C., Ward, P. S., Patel, J., Shih, A., … Melnick, A. (2010). Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell*, *18*(6), 553–67. doi:10.1016/j.ccr.2010.11.015

Floridi, L. (2008). The Method of Levels of Abstraction. *Minds and Machines*, *18*(3), 303–329. doi:10.1007/s11023-008-9113-7

Floridi, L. (2010). *Information: A very short introduction*. Oxford University Press.

Forber, P. (2011). Reconceiving Eliminative Inference. *Philosophy of Science*, *78*(2), 185–208.

Franklin, A. (1999). *Can That Be Right? Essays on Experiment, Evidence, and Science*. Dordrecht: Kluwer Academic Publishers.

Freedman, M. L., Monteiro, A. N. a, Gayther, S. a, Coetzee, G. a, Risch, A., Plass, C., … Mills, I. G. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. *Nature Genetics*, *43*(6), 513–8. doi:10.1038/ng.840

Gannon, F. (2009). A letter to Darwin. *EMBO Reports*, *10*(1), 1. doi:10.1038/embor.2008.239

Garraway, L. a, & Lander, E. S. (2013). Lessons from the cancer genome. *Cell*, *153*(1), 17–37. doi:10.1016/j.cell.2013.03.002

Garraway, L. a, & Lander, E. S. (2013). Lessons from the cancer genome. *Cell*, *153*(1), 17–37. doi:10.1016/j.cell.2013.03.002

Germain, P., Ratti, E., & Boem, F. (2014). Junk or Functional DNA? ENCODE and the Function Controversy. *Biology & Philosophy*, *Forthcomin*.

Giordano, T. J. (2014). The Cancer Genome Atlas Research Network: A Sight to Behold. *Endocrine Pathology*, *25*(4), 362–365. doi:10.1007/s12022-014-9345-4

Goldfarb, M., Shimizu, K., Perucho, M., & M., W. (1982). Isolation and preliminary characterization of a human transforming gene from T24 bladder carcinoma cells. *Nature*, *296*.

Goldfarb, M., Shimizu, K., Perucho, M., & M., W. (1982). Isolation and preliminary characterization of a human transforming gene from T24 bladder carcinoma cells. *Nature*, *296*.

Goldman, A. (1988). Strong and Weak Justification. *Philosophical Perspectives*, *2*, 51–69.

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

Grieb, B. C., Chen, X., & Eischen, C. M. (2014). MTBP is overexpressed in triple-negative breast cancer and contributes to its growth and survival. *Molecular Cancer Research : MCR*, *12*(9), 1216–24. doi:10.1158/1541-7786.MCR-14-0069

Gross, F. (2013). *The Sum of the Parts: Heuristic Strategies in Systems Biology*. University of Milan.

Guessous, I., Gwinn, M., & Khoury, M. J. (2009). Genome-wide association studies in pharmacogenomics: untapped potential for translation. *Genome Medicine*, *1*(4), 46. doi:10.1186/gm46

Guo, W., Keckesova, Z., Donaher, J. L., Shibue, T., Tischler, V., Reinhardt, F., … Weinberg, R. a. (2012). Slug and Sox9 cooperatively determine the mammary stem cell state. *Cell*, *148*(5), 1015–28. doi:10.1016/j.cell.2012.02.008

Hacking, I. (1992). The self-vindication of the laboratory sciences. In A. Pickering (Ed.), *Sciences as Practice and Culture*. Chicago: University of Chicago Press.

Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., … Pericak-vance, M. A. (2005). Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. *Science*, *308*(April), 419–422.

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer Review University of California at San Francisco, *100*, 57–70.

Hanahan, D., & Weinberg, R. a. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–74. doi:10.1016/j.cell.2011.02.013

Hanson, N. (1958). *Patterns of Discovery*. Cambridge, Uk: Cambridge University Press.

Hanson, N. R. (1958). The logic of discovery. *The Journal of Philosophy*, *55*(25).

Hanson, N. R. (1960). Is there a logic of scientific discovery? *Australasian Journal of Philosophy*, *38*(2).

Hansson, S. O. (2007). Values in pure and applied science. *Foundations of Science*, *12*(3), 257–268. doi:10.1007/s10699-007-9107-6

Harman, G. H. (1965). The inference to the best explanation. *Philosophical Review*, *74*.

Harman, G. H. (1968). Enumerative induction as inference to the best explanation. *Journal of Philosophy*, *65*.

Hawthorne, J. (1993). Bayesian induction is eliminative induction. *Philosophical Topics*, *21*(1).

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press.

Hempel, C. (1966). *Philosophy of Natural Science*. Upper Saddle River, New Jersey: Prentice-Hall.

Hertz, H. (1899). *The Principles of Mechanics : Presented in a New Form*. London: Macmillan.

Hood, L., & Rowen, L. (2013). The human genome project: big science transforms biology and medicine. *Genome Medicine*, *5*(9), 79. doi:10.1186/gm483

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., … Yang, H. (2010). International network of cancer genome projects. *Nature*, *464*(7291), 993–8. doi:10.1038/nature08987

Huene, P. H. (2006). Context of discovery versus context of justification and Thomas Kuhn. In J. Schickore & F. Steinle (Eds.), *Revisiting Discovery and Justification* (pp. 119–131). Springer Netherlands.

Hunter, D., Altshuler, D., & Rader, D. (2008). From Darwin's Finches to Canaries in the Coal Mine - Mining the Genome for the New Biology. *The New England Journal of Medicine*.

Hunter, D., Altshuler, D., & Rader, D. (2008). From Darwin's Finches to Canaries in the Coal Mine - Mining the Genome for the New Biology. *The New England Journal of Medicine*.

Hunter, D., Altshuler, D., & Rader, D. (2008). From Darwin's Finches to Canaries in the Coal Mine - Mining the Genome for the New Biology. *The New England Journal of Medicine*.

Kay, L. (2000). *Who wrote the book of life? A History of the Genetic Code*. Stanford: Stanford University Press.

Kay, L. (2000). *Who wrote the book of life? A History of the Genetic Code*. Stanford: Stanford University Press.

Keating, P., & Cambrosio, A. (2012). Too many numbers: Microarrays in clinical cancer research. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 37–51. doi:10.1016/j.shpsc.2011.10.004

Kelly, K. T. (1987). The Logic of Discovery. *Philosophy of Science*, *54*(3), 435–452.

Kim, T.-M., Xi, R., Luquette, L. J., Park, R. W., Johnson, M. D., & Park, P. J. (2013). Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Research*, *23*(2), 217–27. doi:10.1101/gr.140301.112

Kincaid, H., Dupre, J., & Wylie, A. (2007). *Value-Free Science? Ideals and Illusions*. Oxford University Press.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*(4).

Kitcher, P. (1993). *The Advancement of Science*. New York: Oxford University Press.

Kitcher, P. (2001). *Science, Truth and Democracy*. Oxford University Press.

Kitsios, G., & Zintzaras, E. (2009). Genome-Wide Association Studies: hypothesis-free or "engaged"? *Translational Research*, *154*(4), 161–164.

Korbel, J. O., & Campbell, P. J. (2013). Criteria for inference of chromothripsis in cancer genomes. *Cell*, *152*(6), 1226–36. doi:10.1016/j.cell.2013.02.023

Kordig, C. (1978). Discovery and justification. *Philosophy of Science*, *45*(1).

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Kuhn, T. (1977). Objectivity, Value Judgement and Theory Choice. In *The Essential Tension: Selected Studies in the Scientific Tradition and Change* (pp. 356–367). Chicago: University of Chicago Press.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. doi:10.1038/35057062

Langley, P., Simon, H., Bradshaw, G., & Zytkow, J. (1987). *Scientific Discovery - Computational Explorations of the Creative Processes*. Cambridge, MA: The MIT Press.

Laudan, L. (1980). Why was the logic of discovery abandoned? In T. Nickles (Ed.), *Scientific Discovery: Logic and Rationality* (pp. 173–184). Dordrecht: Reidel Publishing Company.

Laudan, L. (1980). Why was the logic of discovery abandoned? In T. Nickles (Ed.), *Scientific Discovery: Logic and Rationality* (pp. 173–184). Dordrecht: Reidel Publishing Company.

Laudan, L. (1984). *Science and Values: The Aims of Science and Their Role in Scientific Debate*. University of California Press.

Laudan, L. (2004). The Epistemic, the Cognitive, and the Social. In P. Machamer & G. Wolters (Eds.), *Science, Values and Objectivty*. University of Pittsburgh Press.

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. doi:10.1038/nature12213

Ledford, H. (2010). The cancer genome challenge. *Nature*, *464*(April).

Leonelli, S. (2009). On the Locality of Data and Claims about Phenomena. *Philosophy of Science*, *76*(December), 737–749.

Leonelli, S. (2011). Packaging Data for Re-use: Databases in Model Organism Biology. In P. Howlett & M. S. Morgan (Eds.), *How Well Do Facts travel? The Dissemination of Reliable Knowledge*. Cambridge, MA: Cambridge University Press.

Leonelli, S. (2012a). Classificatory Theory in Data-intensive Science : The Case of Open Biomedical Ontologies. *International Studies in the Philosophy of Science*, (May), 37–41.

Leonelli, S. (2012b). Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 1–3. doi:10.1016/j.shpsc.2011.10.001

Leonelli, S., & Ankeny, R. a. (2012). Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 29–36. doi:10.1016/j.shpsc.2011.10.003

Levins, R. (1966). The strategy of model building in population biology. In E. Sober (Ed.), *Conceptual Issues in Evolutionary Biology* (pp. 18–27). Cambridge, MA: MIT Press.

Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., … Wilson, R. K. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, *456*(7218), 66–72. doi:10.1038/nature07485

Li, Y., Zhang, L., Ball, R. L., Liang, X., Li, J., Lin, Z., & Liang, H. (2012). Comparative analysis of somatic copy-number alterations across different human cancer types reveals two distinct classes of breakpoint hotspots. *Human Molecular Genetics*, *21*(22), 4957–65. doi:10.1093/hmg/dds340

Liu, J., McCleland, M., Stawiski, E. W., Gnad, F., Mayba, O., Haverty, P. M., … Zhang, Z. (2014). Integrated exome and transcriptome sequencing reveals ZAK isoform usage in gastric cancer. *Nature Communications*, *5*(May), 3830. doi:10.1038/ncomms4830

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Longino, H. (1995). Gender, Politics, and the Theoretical Virtues. *Synthese*, *104*(3), 383–397.

Longino, H. (1996). Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy. In L. H. Nelson & J. Nelson (Eds.), *Feminism, Science, and the Philosophy of Science* (pp. 39–58). Kluwer Academic Publishers.

Lu, C., Ward, P. S., Kapoor, G. S., Rohle, D., Turcan, S., Abdel-Wahab, O., … Thompson, C. B. (2012). IDH mutation impairs histone demethylation and results in a block to cell differentiation. *Nature*, *483*(7390), 474–8. doi:10.1038/nature10860

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

Manolio, T. A., & Collins, F. S. (2009). The HapMap and genome-wide association studies in diagnosis and therapy. *Annu. Rev. Med.*, *60*, 443–456. doi:10.1146/annurev.med.60.061907.093117.The

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

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. a, & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews. Genetics*, *9*(5), 356–69. doi:10.1038/nrg2344

Mcmullin, E. (1983). Values in Science. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *2*, 686–709.

Mcmullin, E. (1985). Galilean idealization. *Studies in the History and Philosophy of Science Part A*, *16*(3).

Mitchell, S. D. (1996). Pragmatic Laws. *Philosophy of Science*, *64*(Supplement Proceedings of the 1996 Biennial Meetings of PSA).

Morange, M. (1998). *A History of Molecular Biology*. Cambridge, Massachusetts, and London, England: Harvard University Press.

Morin, R. D., Mendez-Lago, M., Mungall, A. J., Goya, R., Mungall, K. L., Corbett, R. D., … Marra, M. a. (2011). Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*, *476*(7360), 298–303. doi:10.1038/nature10351

Morrison, M. (2000). *Unifying Scientific Theories: Physical Concepts and Mathematical Structures*. Cambridge University Press.

Newell, A., Shaw, J., & Simon, H. (1962). The processes of creative thinking. In H. Gruger, G. Terrell, & M. Wertheimer (Eds.), *Contemporary Approaches to Creative Thinking*. New York: Atherton Press.

Nickles, T. (1980). Introductory essay: scientific discovery and the future of philosophy of science. In T. Nickles (Ed.), *Scientific Discovery: Logic and Rationality* (pp. 1–60). Dordrecht: Reidel Publishing Company.

Nickles, T. (1985). Beyond Divorce: Current Status of the Discovery Debate. *Philosophy of Science*, *52*(2), 177–206.

Norton, J. (1995). Eliminative induction as a method of discovery: How Einstein discovered general relativity. In J. Leplin (Ed.), *The Creation of Ideas in Physics*. Kluwer Academic Publishers.

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. doi:10.1016/j.shpsc.2011.10.006

O'Malley, M. a, Elliott, K. C., Haufe, C., & Burian, R. M. (2009). Philosophies of funding. *Cell*, *138*(4), 611–5. doi:10.1016/j.cell.2009.08.008

O'Malley, M. a. (2007). Exploratory experimentation and scientific practice: Metagenomics and the proteorhodopsin case. *History and Philosophy of the Life Sciences*, *29*(3), 335–358.

Odenbaugh, J. (2003). Complex systems, trade-offs and mathematical modeling: Richard Levins. "Strategy of model building in population biology" revisited. *Philosophy of Science*, *70*.

Pera, M. (1981). Inductive Method and Scientific Discovery. In M. Grmek, R. Cohen, & G. Cimino (Eds.), *On Scientific discovery*. Dordrecht: Reidel Publishing Company.

Piazza, R., Valletta, S., Winkelmann, N., Redaelli, S., Spinelli, R., Pirola, A., … Gambacorti-Passerini, C. (2013). Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nature Genetics*, *45*(1), 18–24. doi:10.1038/ng.2495

Platt, J. R. (1964). Strong Inference. *Science*, *146* (3642).

Plutynski, A. (2013). Cancer and the goals of integration. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*(4), 466–476. doi:10.1016/j.shpsc.2013.03.019

Pomerantz, M. M., Shrestha, Y., Flavin, R. J., Regan, M. M., Penney, K. L., Mucci, L. a, … Freedman, M. L. (2010). Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis. *PLoS Genetics*, *6*(11), e1001204. doi:10.1371/journal.pgen.1001204

Popper, K. (2002). *The Logic of Scientific Discovery*. London and New York: Routledge.

Pritchard, D. (2007). Recent work on epistemic value. *American Philosophical Quaterly*, *44*(2).

Pulciani, S., Santos, E., Lauver, A. V., Long, L. K., Robbins, K. C., & Barbacid, M. (1982). Oncogenes in human tumor cell lines: molecular cloning of a transforming gene from human bladder carcinoma cells. *Proc. Natl. Acad. Sci. USA*, *79*.

Pulciani, S., Santos, E., Lauver, A. V., Long, L. K., Robbins, K. C., & Barbacid, M. (1982). Oncogenes in human tumor cell lines: molecular cloning of a transforming gene from human bladder carcinoma cells. *Proc. Natl. Acad. Sci. USA*, *79*.

Raphael, B. J., Dobson, J. R., Oesper, L., & Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Medicine*, *6*(1), 5. doi:10.1186/gm524

Ratti, E. (2014). Levels of abstraction, emergentism and artificial life. *Journal of Experimental & Theoretical Artificial Intelligence*, *27*(1), 51–61. doi:10.1080/0952813X.2014.940144

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

Reichenbach, H. (1951). *The Rise of Scientific Philosophy*. University of California Press.

Reichenbach, H. (1961). *Experience and Prediction* (Phoenix Ed.). The University of Chicago Press.

Rescher, N. (1958). On Prediction and Explanation. *The British Journal for the Philosophy of Science*, *8*(32), 281–290.

Rheinberger, H.-J. (1997a). Experimental complexity in biology: Some epistemological and historical remarks. *Philosophy of Science*, *64*(4).

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

Rheinberger, H.-J. (2007). What happened to molecular biology? *B.I.F. Futura*, *22*, 218–223.

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

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., … Sullivan, P. F. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, *45*(10), 1150–9. doi:10.1038/ng.2742

Rooney, P. (1992). On Values in Science : Is the Epistemic/Non-Epistemic Distinction Useful? In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (Vol. 1992, pp. 13–22).

Rudner, R. (1953). The scientist qua scientist makes value judgement. *Philosophy of Science*, *20*(1), 1–6.

Russell, N. (1959). On the Symmetry Between Explanation and Prediction. *The Philosophical Review*, *68*(3), 349–358.

Salmon, W. (1967). *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.

Salmon, W. (1978). Why ask, "why?"? An inquiry concerning scientific explanation. In *Proceedings and Adressing of the American Philosophical Association* (pp. 683–705).

Santarius, T., Shipley, J., Brewer, D., Stratton, M. R., & Cooper, C. S. (2010). A census of amplified and overexpressed human cancer genes. *Nature Reviews. Cancer*, *10*(1), 59–64. doi:10.1038/nrc2771

Schaffner, K. (1993). *Discovery and Explanation in Biology and Medicine*. Chicago: The University of Chicago Press.

Schaub, M. a, Boyle, A. P., Kundaje, A., Batzoglou, S., & Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Research*, *22*(9), 1748–59. doi:10.1101/gr.136127.111

Schickore, J. (2014). Scientific Discovery. In *The Stanford Encyclopedia of Philosophy*.

Scriven, M. (1959). Explanation and Prediction in Evolutionary Theory. *Science*, *130*.

Shih, C., & Weinberg, R. (1982). Isolation of a transforming sequence from a human bladder carcinoma cell line. *Cell*, *29*.

Shih, C., & Weinberg, R. (1982). Isolation of a transforming sequence from a human bladder carcinoma cell line. *Cell*, *29*.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. doi:10.1214/10-STS330

Simon, H. (1977). *Models of Discovery*. Reidel: Dordrecht.

Smalheiser, N. R. (2002). Informatics and hypothesis-driven research. *EMBO Reports*, *3*(8), 702. doi:10.1093/embo-reports/kvf164

Stehelin, D., Varmus, H. E., Bishop, J. M., & Vogt, P. K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, *260*.

Stehelin, D., Varmus, H. E., Bishop, J. M., & Vogt, P. K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, *260*.

Steinle, F. (1997). Entering New Fields: Exploratory Uses of Experimentation. *Philosophy of Science*, *64*(Supplement. Proceedings of the 1996 Biennial Meetings of PSA), 64–74.

Stransky, N., Egloff, A. M., Tward, A. D., Kostic, A. D., Cibulskis, K., Sivachenko, A., … Grandis, J. R. (2011). The mutational landscape of head and neck squamous cell carcinoma. *Science (New York, N.Y.)*, *333*(6046), 1157–60. doi:10.1126/science.1208130

Strasser, B. (2011). The Experimenter ’s Museum - GenBank , Natural History , and the Moral Economies of Biomedicine. *Isis*, *102*(1), 60–96.

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

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., … Lopez-Bigas, N. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports*, *3*, 2650. doi:10.1038/srep02650

Teller, P. (2010). "Saving the Phenomena" Today. *Philosophy of Science*, *77*(5), 815–826.

Thagard, P. (1982). Artificial Intelligence, Psychology, and the Philosophy of Discovery. In *PSA : Proceedings of the Biennial Meeting of the Philosophy of Science Association , Vol . 1982, Volume Two : Symposia and Invited Papers (1982)*.

The Cancer Genome Atlas Consortium. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, *507*(7492), 315–22. doi:10.1038/nature12965

The ENCODE Consortium. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, *306*(5696), 636–640. doi:10.1126/science.1105136

The ENCODE Project Consortium. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology*, *9*(4), e1001046. doi:10.1371/journal.pbio.1001046

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi:10.1038/nature11247

The Gene Ontology Consortium. (2000). Gene Ontology : tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. doi:10.1038/75556.Gene

Van Fraassen, B. (1980). *The Scientific Image*. Oxford University Press.

Vandin, F., Upfal, E., & Raphael, B. J. (2011). Finding driver pathways in cancer: models and algorithms. *Algorithms for Molecular Biology : AMB*, *7*(1), 23. doi:10.1186/1748-7188-7-23

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, *291*(5507), 1304–51. doi:10.1126/science.1058040

Visscher, P. M., Brown, M. a, McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24. doi:10.1016/j.ajhg.2011.11.029

Vogelstein, B., Fearon, E. R., Kern, S. E., S.R., H., Preisinger, A. C., Nakamura, Y., & White, R. (1989). Allelotype of colorectal carcinomas. *Science*, *244*.

Vogelstein, B., Fearon, E. R., Kern, S. E., S.R., H., Preisinger, A. C., Nakamura, Y., & White, R. (1989). Allelotype of colorectal carcinomas. *Science*, *244*.

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. doi:10.1126/science.1235122

Waters, C. K. (2007). The Nature and Context of Exploratory Experimentation. *History and Philosophy of the Life Sciences*, *29*, 1–9.

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

Weinberg, R. a. (2014). Coming full circle-from endless complexity to simplicity and back again. *Cell*, *157*(1), 267–71. doi:10.1016/j.cell.2014.03.004

Weinberg, R. a. (2014). Coming full circle-from endless complexity to simplicity and back again. *Cell*, *157*(1), 267–71. doi:10.1016/j.cell.2014.03.004

Weisberg, M. (2004). Qualitative theory and chemical explanation. *Philosophy of Science*, *71*.

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

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Whitt, L. A. (1990). Theory pursuit: Between discovery and acceptance. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*.

Wimsatt, W. (2007). *Re-engineering Philosophy for Limited Beings - Piecewise Approximations to Reality*. Cambridge, Massachusetts, and London, England: Harvard University Press.

Woodward, J. (2000). Data , Phenomena , and Reliability. *Philosophy of Science*, *67*.

Woodward, J. F. (2011). Data and phenomena: a restatement and defense. *Synthese*, *182*(1), 165–179. doi:10.1007/s11229-009-9618-5

Woodward, J. I. M. (1989). Data and Phenomena. *Synthese*, *79*(3), 393–472.

Yaffe, M. B. (2013). The Scientific Drunk and the Lamppost. *Science Signaling*, *6*(269), 1–3.

Yeo, G. S. H. (2011). Where next for GWAS? *Briefings in Functional Genomics*, *10*(2), 51. doi:10.1093/bfgp/elr011

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., … Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, *45*(10), 1134–1140. doi:10.1038/ng.2760

Zamecnik, P. C. (1962). History and speculation on protein synthesis. *Proceedings of the Symposia on Mathematical Problems in the Biological Sciences*, *14*.

Zang, Z. J., Cutcutache, I., Poon, S. L., Zhang, S. L., McPherson, J. R., Tao, J., … Tan, P. (2012). Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nature Genetics*, *44*(5), 570–4. doi:10.1038/ng.2246