LARGE-SCALE BIOLOGY: PHILOSOPHICAL, HISTORICAL, AND COMPUTATIONAL PERSPECTIVES

Emanuele Ratti[1]

Johannes Kepler University Linz

Thomas Stoeger,

Northwestern University

Abstract. The relation between the ethos of large-scale projects in the life sciences and the epistemic culture of molecular biology has been the subject of heated discussions for the past 30 years. Molecular biology is typically a 'small science', organized around a laboratory leader who decides what to pursue, placing 'bets' on different research strategies. 'Large-scale' biology has taken several forms, from centralized computational infrastructures, to big consortia distributed throughout a country and distributing efforts among many 'small' units. While several scholars have analyzed how large-scale biology has impacted the sociological, epistemic, and the governance structure of molecular biology, works investigating how its discovery strategies have been shaped by large-scale projects have been elusive. In this paper, we identify two ways in which large-scale biology could possibly influence the discovery strategies of traditional molecular biology. First, large-scale projects can be facilitators: while small labs continue to pursue their own interests, the way their hypotheses are developed is facilitated by the resources provided by large-scale projects. Second, large-scale projects can be epistemically centralizing: they set the research agenda of molecular biology labs by constraining the choices of hypotheses to pursue in the first place. As an example of the former, we discuss the Human Genome Project from a historical and epistemic angle. As an example of the latter, we discuss The Cancer Genome Atlas. In order to explore the effects of this project on discovery strategies, we perform a thorough computational analysis of the literature on cancer gene from the 1980s to 2018.

1. INTRODUCTION

---

[1] mnl.ratti@gmail.com

In her seminal study of the knowledge culture of molecular biology, Knorr-Cetina (1999) characterizes molecular labs as a "benchwork science conducted in small laboratories" (p 81). Small labs in molecular biology, in her recollection, are organized around a laboratory leader who decides what to pursue and what not to pursue, placing 'bets' on different research trajectories. When setting up a new laboratory, a laboratory leader organizes it according to the particular tradition he/she comes from, in the sense that research questions, methodologies and approaches are built on his/her previous experience. Starting from the 1980s, this representation of molecular biology as 'small-lab' knowledge culture has been shaken by the emergence of large-scale projects in the life sciences. 'Large-scale' has taken several forms in the experimental life sciences, ranging from centralized computational infrastructures (such as GenBank), to big consortia distributed throughout a country and distributing efforts among many 'small' units[2] (such as The Cancer Genome Atlas or ENCODE). Contrasts have been drawn between these two ways of organizing the experimental life sciences. For instance, large-scale projects, by being 'big science', have been seen as a threat by molecular biologists, whose culture reflects a small-scale type of science (Morange 2020), which takes pride on a relative 'autonomy' of laboratory leaders to pursue their own projects. Another concern is that large-scale projects have been seen as diverting resources away from small labs (Weinberg 1987), thereby making difficult for laboratory leaders to pursue their projects. Moreover, some say that large-scale projects such as Human Genome Project (HGP) or ENCODE reflect a 'collecting' way of knowing that is not scientifically valuable for molecular biologists, who see experiments as the only access to the secrets of nature (Weinberg 2010; Alberts 2012; Strasser 2019) and to the 'empirical' (Knorr-Cetina 1999). For all these reasons, some molecular biologists have been suspicious of the impact of large-scale biology, and sometimes even threatened by it.

Studies of large-scale biology and their impact on experimental life sciences are not scarce (Stevens 2013; Richardson and Stevens 2015; Leonelli 2016; Hilgartner 2017; Strasser 2019). Several scholars have analyzed how large-scale biology has impacted the sociological, epistemic, and the governance structure of molecular biology which is typically, as already noted, a science organized around small and autonomous laboratories[3]. In particular, studies have explored how

---

[2] We are indebted to Christopher Donohue for this observation
[3] We will get into more detail later, but right now 'small' and 'autonomous' is defined along the lines of Knorr-Cetina's considerations mentioned at the beginning

centralized collecting efforts typical of large-scale projects have modified the system of rewards of the life sciences (Strasser 2019), have created new professional roles (Leonelli 2016), generated new theoretical tools (Leonelli 2016) and even new disciplines (Stevens 2013; Strasser 2019). But issues regarding how big biology has affected and constrained the discovery strategies of molecular biological small labs have not been fully addressed. In this paper, we analyze how large-scale biology has influenced and could possibly influence the epistemic strategies of traditional molecular biology. By 'epistemic strategy' and 'discovery strategy' we mean how hypotheses are generated and developed. In Section 2 we discuss more in detail what a discovery strategy is in molecular biology by elaborating an account in which such strategies are composed of three phases. By applying this account in Section 3 and 4, we discuss how two large-scale projects (HGP and TCGA) have impacted the discovery strategies of molecular biology. We discuss these two projects in particular because they exemplify (at least potentially) two prominent ways in which large-scale biological projects could impact the epistemic strategies of 'small-scale' molecular biology.

First, large-scale projects can facilitate the everyday work of small and autonomous biological labs. While small labs continue to pursue their own interests, the way their hypotheses are developed is facilitated by the resources provided by large-scale projects. We claim that such large-scale projects are *facilitators or accelerators* because they aid 'small biology' by accelerating the ways in which biologists develop research questions and hypotheses which are formulated independently. We will show how the HGP fits exactly this bill, despite controversies ascribing to it a much more disruptive role.

The second type of impact is more robust. Rather than just providing useful tools to develop hypotheses in a more efficient way, large-scale projects may also constraint the choice of hypotheses in the first place, by providing reasons to pursue certain hypotheses rather than others. In this way, even the starting point of research of small labs is constrained by large-scale projects. In other words, large-scale biology, by providing justificatory reasons to pursue certain hypotheses rather than others, can *potentially establish—imperialistically—the research agenda of small biological labs*. This type of impact also can have, in the long run, effects on the governance of molecular biology, the way it is funded, and foster the dependence of small biological labs on large-scale projects. We say that these projects are *epistemically centralizing* because, by dictating research questions for small labs, they centralize the epistemic dimension of the discipline.

However, whether a project with such an impact exists is an open question. The Cancer Genome Atlas (TCGA) is a candidate project, and we discuss this possibility. In particular, given the particular epistemic structure of cancer genomics, we should see a significant growth in research on new cancer genes discovered by TCGA after 2008 (the year of the first cancer genome study released by TCGA). This would show that biologists have since the inception of TCGA explored the biological properties and mechanisms of the genes uncovered by TCGA, and hence their research agendas have been dictated by TCGA itself. However, we perform a computational analysis of literature on cancer genes up through 2018 (Repana et al 2019) and found that this did not happen and that most cancer genes known today have been discovered and studied before TCGA. But even if TCGA does not fit the characteristics of large-scale projects that impact biology in the second sense, still our hypothesis captures a trend in the governance of small-scale biology. Agencies such as the National Institutes of Health have issued calls for funding that are explicitly directed towards the use of data from large-scale projects.[4]

2 STRATEGIES OF DISCOVERY IN MOLECULAR BIOLOGY

2.1 A primer on discovery strategies – A tripartite account

In order to understand how large-scale projects in biology influence discovery strategies and whether they may possibly undermine the epistemic independence of molecular biology labs, we should say something about discovery strategies in general. It is commonplace to distinguish between a phase where a hypothesis or model is generated and a phase where evidence is provided in support of a hypothesis or models are evaluated for accuracy according to certain standards.[5] However, some philosophers have realized that such a sharp division does not do justice to the variety of processes that hypotheses are subjected to (Nickles 1980; Curd 1980; Schaffner 1993). These scholars make a distinction between *hypothesis formation/generation* and *hypothesis weak evaluation/development* in order to show how the discovery phase is much more complicated

---

[4]See for instance https://grants.nih.gov/grants/guide/rfa-files/rfa-rm-18-021.html,
https://grants.nih.gov/grants/guide/rfa-files/RFA-RM-19-012.html

[5] We are not referring to the distinction between the context of discovery and the context of justification, especially not in the terms originally formulated by Reichenbach in *Experience and Prediction* (1961). Rather, we are focusing on what happens typically in the context of discovery, thereby assuming – *contra* Reichenbach – that in the context of discovery it is possible to sharply identify aspects of scientific practice that are epistemically relevant.

(Curd 1980). According to Curd, 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, we can distinguish three broad, procedural phases of scientific practice: hypothesis generation, hypothesis prior assessment/development, and hypothesis justification/confirmation. The idea behind the phase of prior assessment/hypothesis development is that theories or hypotheses are not generated in the final form in which they are tested (the phase of justification). Theories/hypotheses are subjected to a 'period of incubation' (Duhem 1954, p 221). In this period of incubation, hypotheses and theories are developed (i.e., the second phase).

Abstracting from the complexities of scientific practice, we can think of discovery strategies as being composed by these three broad phases: hypothesis generation, hypothesis development, and hypothesis confirmation. During the phase of hypothesis generation, scientists construct or build a preliminary hypothesis or model. Such a model will be generated in many different ways, such as from intuitions or by building on previous knowledge. The process of generating a hypothesis has been conceived by some philosophers as irrational (or arational) because it is sometimes opaque. The phase of hypothesis development is when the model or hypothesis generated is subjected to modifications because of actual scientific work. While the phase of hypothesis development is often standardized in each scientific discipline according to norms for experiments and topical techniques (inter alia), the phase of hypothesis generation may be seemingly random even within the same discipline. Finally, the phase of confirmation is when a certain model or hypothesis, after it has undergone a great deal of modifications and refinements, is finally considered adequate and detailed enough to be tested.[6] This characterization should not be understood as providing an abstract structure to impose on scientific practice, nor does it mean that any discovery episode can be characterized sharply in terms of "first comes generation, then development and finally confirmation." The process involves reciprocal feedback and it is iterative.

2.2 The tripartite account in molecular biology
This tripartite characterization of scientific discovery makes it possible to figure out the contribution of large-scale tools to each phase of scientific discovery and thereby understand the

---

[6] Sometimes it is not clear what kinds of operations are involved in this phase of 'strong' evaluation.

epistemic impact of large-scale biology on molecular biology. Discovery strategies in molecular biology are intimately connected to experimentation, and especially material manipulation. Molecular biologists gather information about phenomena by modifying them, and to modify them one needs to develop standardized techniques to constrain and exploit the internal dynamics of cells, as well as relations between cells.

Molecular biologists are constrained by an experimental system, which is defined as the "integral, locally manageable, functional unit of scientific research" (Rheinberger 1997b, p S246). An experimental system may be conceived as the particular phenomenon one is investigating, together with the conceptual and experimental apparatus used to grasp its dynamics. Researchers are tightly committed to their selected experimental system:

"In analyzing a problem, the biologist is constrained to focus on a fragment of reality, on a piece of the universe which he arbitrarily isolates to define certain of its parameters. In biology, any study thus begins with the choice of a 'system'" (Jacob 1988, p 234)

After the experimental system is chosen, biologists 'attack' the epistemic complexity of biological phenomena (i.e., what they have limited access to) by constraining and manipulating the experimental system (both conceptually and materially) in various ways. For this reason, Rheinberger focuses his analysis on the details of epistemic and experimental practices.

Highly complex mechanistic models do not come out of nowhere (i.e., hypothesis generation) and then are tested (i.e. hypothesis justification). In molecular biology something is established via a *system of experiments*. Systems of experiments include several types of experiments, from exploratory (Franklin 2005; Waters 2007) to clear-cut theory-driven ones. Even if the system of experiments is complex, we can still identify at least two of the three phases,[7] though their relations are quite intricate.

As shown in (Boem and Ratti 2016), the discovery process and systems of experiments in molecular biology comprise a *sui generis*, ramified, and intricate hypothetico-deductive system (HD-system). Systems of experiments in molecular biology usually develop and shape hypotheses about the way a particular phenomenon is produced and maintained. There is an emphasis on

---

[7] Hypothesis justification in molecular biology is complicated and set aside here out of necessity.

causal narratives,[8] which is typical of mechanistic models (Glennan 2017). The flow of reasoning is intricate, but we can still use the HD-system to make sense of it. The first step is a general guess about a biological system, such as how it is produced or maintained. This is the phase of hypothesis generation. Here the experimental system is decomposed into entities and activities that are deemed important for the production or maintenance of the phenomenon under scrutiny (Bechtel and Richardson 2010). This is when we "choose" the experimental system and have expectations from it based on previous research, intuitions and other items that philosophers of science in the past have labelled as psychological, sociological, or historical. In molecular biology, this is when the laboratory leader, to use Knorr-Cetina's expression, *place bets* for future research. However, because these guesses are so broad and general, several and often contrasting predictions may be derived. In order to ascertain which predictions are in accordance with the nature of phenomena, one manipulates the experimental system *via* experiments (*à la* Hacking, 1983) in order to stimulate it to 'reveal' more information. Some initial predictions are discarded, while others are transformed into more precise hypotheses. Some predictions "contribute – through additional information they generate – to shape and refine the hypotheses about the functioning of the experimental systems" (Boem and Ratti 2016, p 150). Next, other experiments are repeated to observe reactions of the experimental systems. Then, some hypotheses are further developed and others are discarded. This process continues, virtually, *ad libitum*; "[t]his is a sort of progressive and ramified (but not-linear) deductive process, developed by poking and prodding experimental systems" (p 150).

Therefore, hypothesis generation is the first step towards an intricate composition of experiments leading to a final model. Hypothesis development is the phase that contributes the most to mechanistic models, and it is when systems of experiments are instantiated to develop a hypothesis. One may argue that generation and development are nearly equivalent. However, during hypothesis generation one need not provide strong evidential reasons for a hypothesis or commit to exploring it. It is striking how little evidence about molecular details is required—

---

[8]See for instance Leng and Leng, who argue in (2020) causal narratives lend themselves good for persuading others. Moreover, many articles in molecular biology tell a story, thereby emphasizing the importance of narratives. When one of the authors (E.R.) was a PhD student at the European School of Molecular Medicine (SEMM) in Milan, he attended a course called *Scientific Writing* where the instructor emphasized how important it is to tell a coherent story where each molecular biological experiment plays a key role in the narrative in preparing the next.

details are what biologists look for later on in the process. For instance, if previous studies had noticed that a protein, if mutated, disrupts a particular molecular pathway, one might generate a hypothesis simply from this unconstrained observation. However, neither the actual activity of the protein nor the specific way it is mutated (or even how the pathway was disrupted) have been uncovered in detail. This is the gap that biologists try to fill after generating a very general hypothesis about the way the protein is important to make sense of the phenomenon observed. The first guess is the most important – the choice of strategy, such as decomposition/localization (Bechtel and Richardson 2010) or backward/forward chaining and schema instantiation (Craver and Darden 2013), will depend on it. Next, because of the nature of the system and the techniques at our disposal, we modify and manipulate the experimental system accordingly, to stimulate it to reveal more until our initial guess can become more and more precise—until we have an adequately detailed mechanistic description of what the protein does. In the phase of hypothesis development, some reasons must be provided in order to pursue and develop one specific hypothesis rather than another.

2.3 Using the account to explore the epistemic dimension between small-labs and large-scale projects

Having laid out this general picture of how discovery strategies work in molecular biology, we now can investigate how large-scale biology had an impact on the epistemic strategies of small-scale molecular biological labs. Note that we are not exclusively interested in the way knowledge generated by molecular biologists is influenced and constrained by large-scale biology. If we only have the typical notion of epistemic dependence from analytic epistemology (Hardwig 1985), then the knowledge generated by molecular biologists is clearly but uninformatively dependent on the data generated by large-scale projects. Although the natural sciences exemplify this common phenomenon, here we are interested in the extent to which large-scale biology projects *dictate* to small labs in molecular biology what to do in the three phases of discovery, and thereby constrain the knowledge that they can generate. In other words, we are interested in how large-scale biology channeled the way molecular biological knowledge is generated and developed.

Our hypothesis is that there are two broad possibilities based on the way we have decomposed the context of discovery of molecular biology. First, large-scale biology might *facilitate* or *accelerate* the work of autonomous small-scale biology labs, for instance by providing

resources that are used within the phase of hypothesis development. This is the phenomenon of standardization that yields experimental shortcuts. We call these large-scale project *facilitators*. In this scenario, we claim that labs retain a relative epistemic independence and autonomy, because they are free to choose the starting point of their research (i.e. hypothesis generation). Second, large-scale consortia can provide, in principle, resources not just to pursue and develop hypotheses that small labs independently generate, but rather they *provide the hypotheses themselves*, together *with justificatory reasons to pursue them*. This means that small independent labs are not really free to pursue their own hypotheses; there is an (epistemic) authority that decides what is and is not worth pursuing. Ideally, such consortia explicitly provide *justificatory reasons* to pursue certain hypotheses rather than others. We say that these consortia are *epistemically centralizing*, because they tend to dictate the very starting point of research strategies.[9]

To explore these possibilities, we discuss the HGP (Section 3), which, in our opinion, exemplifies *large-scale facilitator projects*. In Section 4, we hypothesize that TCGA is an example of an epistemically centralizing project and evaluate the merit of this conjecture by means of computational analysis.

3 IMPACTS OF HGP ON MOLECULAR BIOLOGY DISCOVERY STRATEGIES

In (1986), Renato Dulbecco advocated for the idea of a sequencing project of the human genome. In commenting on different approaches to molecular research in cancer, he added:

"I think that it will be far more useful to begin by sequencing the cellular genome. The sequence will make it possible to prepare probes for all the genes and to classify them for their expression in various cell types at the level of individual cells by means of cytological hybridization. The classification of the genes will facilitate the identification of those involved in progression" (p 1055)

---

[9] We recognize that a third possibility can arise in the context of 'systems biology', where studies from large-scale laboratories support small-scale laboratories to pursue hypothesis that integrate knowledge about many genes. In such a scenario, large-scale studies gathering measurements facilitate and accelerate the analytical work of autonomous small-scale laboratories, while simultaneously providing the hypothesis that the interplay of many distinct genes – rather than the characterization of individual genes – will yield biological insight. The intertwined relationship between systems biology and genomics – including the Human Genome Project and TCGA – has been excellently reviewed elsewhere (Conesa and Mortazavi 2014; Westerhoff and Palsson 2004)

Dulbecco was not alone in pushing for a large-scale project aimed at sequencing whole genomes (Morange 2020). In 1985, Robert Sinsheimer (chancellor of UC Santa Cruz) organized a conference with the explicit aim of discussing a large-scale project to sequence the human genome. Charles DeLisi of the Department of Energy (DOE) proposed the sequencing of the human genome as a long-term project. Walter Gilbert was collecting opinions about the feasibility of this project in various meetings at the Cold Spring Harbor Laboratory. Therefore, during the second half of the 1980s, what Hilgartner (2017) calls a 'sociotechnical vanguard' emerged to propose what would be later called the HGP. It is important to stress the strategic nature of these individuals: "[l]eaders of sociotechnical vanguards typically assume a visionary role, performing the identity of one who possesses superior knowledge of an emerging development and aspires to realize its more desirable potentials" (Hilgartner 2017, p 27). While the initial vision was on sequencing, it shifted towards a series of mapping projects, then again on pilot projects to develop the required sequencing technology, until the final full-scale sequencing project[10]. DOE decided to grant support for a pilot project in 1987, while one year later NIH envisioned a specific office to work on the project[11] (National Human Genome Research Institute – NHGRI, officially created in 1989). James Watson was put in charge and he was succeeded by Francis Collins in 1992, after he resigned due to a controversy over a conflict of interests about patents (Roberts 1992). Officially, the project started on October 1, 1990, and it was planned to last for a total of 15 years, implemented in five-year plans. HGP grew into an international collaboration which included research projects in France, United Kingdom (with the creation of the Sanger Center) and Japan among the many countries involved (Mueller-Wille and Rheinberger 2012). In parallel to the human genome, genomes of model organisms were sequenced (including *S. Cerevisiae, E. Coli, and D. melanogaster*) in order to develop and calibrate different methodologies. Famously, Venter left the NIH and in 1998 he established Celera Genomics for his own sequencing project. In June 2000, both Venter and Collins announced the completion of drafts of human genome sequence, which were both published in 2001 in *Nature* and *Science*.

In order to understand the initial vision of this project and connect it to our inquiry, we have to take a step back and specify the 1980s 'paradigm' of molecular biology, which remains

---

[10] As per Hilgartner's recollection (2017), the phases are more blurred than my brief and simplistic reconstruction of these events. See also (Hood and Rowen 2013)

[11] For detailed accounts of how this happened see (Cook-Degan 1994) and (Kevles 1997)

similar to what we have today. Molecular biology's main strategy of decomposition and localization was (and *still is*) gene-centered[12]; even though scientists realized—now and then—that DNA is just a small part of cellular complexity,[13] attention focused on DNA as the starting point for investigation[14]. Therefore, if we knew where genes were and what they did, biological understanding would be advanced and research would be greatly facilitated. This is not just a matter of knowledge; experimental techniques are mostly calibrated around DNA (e.g. PCR, microarrays, gene-editing etc). The initial vision of the HGP was to revolutionize molecular genetics - a discipline based on 'gene-hunting' – from this point of view.

3.1 Criticisms of HGP

Many biologists were fiercely opposed to the HGP. This was because of a variety of reasons.

First, the involvement of DOE looked suspicious to many biologists, who did not want to collaborate with physicists who worked for large-scale military projects in the past as they did not want HGP to become a "program for unemployed bomb-makers" (Mueller-Wille and Rheinberger 2012, p 198).

A second reason was that biologists were unease with the proposed scale of the project. The project was interpreted as an attempt to turn biology into a 'big science' discipline as physics, and this was seen "as a threat to their discipline" (Morange 2020, p 342). As Strasser reports (2019), molecular biology was "little science *par excellence*" (p 195), and even if experimental biologists had participated in other centralized and collective efforts before HGP (such as *GenBank* or the *Atlas of Protein Sequence and Structure*), their 'ethos' (to borrow Strasser's terminology) was still individualistic. In order to get an idea of this attitude, Strasser lists the concerns raised during a meeting at Rockefeller University in 1979 to discuss the creation of a centralized sequence database. Among the others, molecular biologists were concerned that a centralized facility would resemble too much postwar-big-science-physics facilities, which was in direct opposition with

---

[12] Waters articulated this idea as 'retooling of genetics', which is the fact that "[g]enes are used as levers to manipulate and investigate a wide variety of biological processes" (2008, p 261). We are indebted to Alan Love for this suggestion

[13] In particular, the complexity of the protein domain seems overwhelming; if we think about the genome as a code, and DNA as its alphabet, we have a language composed of four letters; in the case of proteins we have a language composed of *at least* 20 letters (i.e. amino acids) and that is only for the primary structure

[14] This is true even though attention and resources have been diverted towards non-coding regions and their regulatory roles (I am indebted to Christopher Donohue for this observation)

them "taking pride in the smaller scale of their laboratories" (p 201). While this episode is not about HGP, it interestingly sheds light on the ethos of molecular biologists. It is also interesting to note, as Mueller-Wille and Rheinberger rightly point out (2012), that big science in biology turned out to be quite different from big science projects in physics, "which were typically organized around a very large scientific instrument" (p 201). Unlike in physics, big science in biology developed with 'patchwork' and distributed character around different 'genome centers' (Hilgartner 2001). But independently of the nature of big science in biology, HGP seemed to subtract resources from small, independent labs, which were seen as the real engines of molecular biology. For instance, Davis and colleagues said that "human genome initiative is competing with the initiatives of investigators in other areas" (1990, p 342). Rechsteiner explicitly noted that "HGP diverts funding from university laboratories to centers, thus having a negative impact on training future scientists" (1991, p 455). Weinberg added that this could demotivate investigators from more 'traditional' style of research: "[r]unning laboratories focused on small-scale, hypothesis-driven research has become unattractive for many young people" (Weinberg 2010, p 678).

A third (and more related to the point we want to make) set of criticisms to HGP has to do with the value of the project itself. For instance, Weinberg suggested that sequencing the genome would not answer any particularly interesting question about human biology: "we may understand less about ourselves at the end of this project than when we began" (Weinberg 1987, https://www.the-scientist.com/opinion-old/the-case-against-gene-sequencing-63318 accessed 10/18/2018). Davis and colleagues were more precise and suggested that it is hard "to see how a complete sequence could be useful for understanding the organization of the huge human genome" (p 343). Critics saw HGP as "mediocre science" (Rechsteiner 1991, p 457) for similar reasons. Moreover, the criticisms concerning the value of HGP reproduced an old and well-known opposition between experimental and collecting efforts. Strasser (2019) characterizes this as an opposition between two 'ways of knowing', and while he convincingly shows that the two ways of knowing have been successfully integrated, still old and recent actors used the opposition as a powerful rhetorical tool. For instance, Rechsteiner said that "the finest science is characterized by (…) elegantly designed and competently performed experiments (…) HGP is little more than a massive data collecting effort" (1991, p 459). As Strasser concisely puts it, collecting and comparing "were often regarded as archaic by experimental biologists" (2019, p 214). Even in 2012, Alberts raised similar concerns for another large-scale project (ENCODE): "the grand

challenges that remain in attaining a deep understanding of the chemistry of life will require going beyond detailed catalogs" (2012, p 1583). The epistemic goals of big science and small science were seen as in tension, rather than complementary.

All these issues (involvement of DOE, concerns about big science and centralized efforts, and criticisms of the scientific value of the project) are eminently cultural, and they affect the perception of the epistemic dimension of HGP. As Knorr-Cetina (1999) shows, the knowledge culture in molecular biology is framed in the epistemic and political authority of the laboratory head, who makes each small lab (somehow) epistemically independent to place bets on research projects as he/she sees fit. Even though the HGP did not turn out to be as centralized as large-scale physics, the attempt to make biology a big science was interpreted as a threat to the independence of small labs. To use the terminology introduced above, some molecular biologists were worried that HGP would become an *epistemically centralizing* project. In other words, the fear was that DOE or the NIH would set research goals or long-term plans that eventually would involve small labs, threatening their independence by forcing them to follow the lead of large-scale biology. To frame this concerns in terms of our account of discovery in biology, the question is: in which phase of discovery HGP had the major impact? Did it dictate research agendas of small labs by imposing itself in the phase of hypothesis generation (i.e. an epistemically centralizing project), or did it just become one tool among the many in the phase of hypothesis development (i.e. a facilitator project)?

3.2 Real Impact of HGP

The HGP did not, in fact, undermined small-labs independence in the way some envisioned. Instead of setting new priorities or research agendas, it facilitated the old gene-hunting agenda of 'normal' molecular biology – in other words, it has been a facilitator rather than a centralizer. The HGP merely facilitated and made more efficient what small labs were pursuing independently. Simple examples which reflect common activities in laboratories are easy to find. Having an infrastructure that provides you with the entire sequence of a genome can help you in designing better primers when doing a PCR. "[W]hole-genome sequences are now available for many species. Thus, it is possible in principle to electronically estimate the likely specificity of candidate primer pairs, and one common approach is to search for and then avoid segments of the intended amplification target that are too similar to other sequences" (Untergasser et al 2012, p 2). Having

a complete genome sequence aids researchers in designing primers that are more specific to target sequences. A similar example is provided by microarrays, which are instruments to measure the expression of thousands of genes. Microarrays are microscope slides printed with thousands of spots in specific positions. There is a probe attached to each spot that is a sequence of 50-100 base pairs complementary to a selected mRNA. After collecting mRNA molecules from samples, you discover whether they attach via complementary base-pairing to the probe. But microarrays may have tens of thousands of probes; to design them you need to have as much sequence as possible to avoid redundancies and imprecision, which can be done if you have the entire genome sequence.

These and other examples show that the resources provided by the HGP did not undermine the authority and independence of small molecular biology labs. Instead, they facilitated their work. In particular, the HGP offers resources that facilitate *hypothesis development*, by making existing experimental techniques faster and more efficient. Rather than wasting time in preparing your experimental apparatus, the HGP can help biologists in cutting corners (in a good way). The HGP helps molecular biological research be faster and more efficient where biologists want and need it. Rather than using a trial-and-error approach to design primers or probes, the HGP helps you to avoid redundancies. These experimental shortcuts are incredibly useful in the phase of hypothesis development. From this point of view, HGP transform long and painful processes into routine procedures.

But these points were already clear to acute observers back in the 1980s and 1990s. A famous case is Walter Gilbert who, in a piece entitled "Towards a paradigm shift in biology" (1991), contrasted two ways of attacking problems in biology. The first is to identify a gene by using various techniques, and then try to understand its function by experimental means. But with HGP, all genes 'will be known' and "the starting point of a biological investigation will be theoretical [i.e. a theoretical conjecture]" (p 99). Therefore, biologists will continue to choose their own starting point and pursue their own interests, and HGP will make this process more efficient:

"The actual biology will continue to be done as 'small science' – depending on individual insight and inspiration to produce new knowledge – but the reagents that the scientist will use will include a knowledge of the primary sequence of the organism, together with a list of all previous deductions from that sequence" (p 99)

To use our own terminology, hypotheses will be generated independently by small labs, but the way they are developed will be aided by HGP.

Therefore, a project like the HGP did not reshape the priorities of molecular biology or put small laboratories in a position of epistemic inferiority. Instead, it helped them achieve whatever they aimed to achieve. The resources of the HGP (and other large-scale projects in biology) take the form of a map: "a data resource comprehensive, complete, closed-ended – to be used my multiple groups, over a long time, for multiple purposes" (Eddy 2013, p 261). The map makes it possible to navigate genome complexity in order to better understand cellular complexity.

4 THE CANCER GENOME ATLAS: AN EPISTEMICALLY CENTRALIZING EFFORT?
If concerns about the impact of the HGP on the epistemic independence of molecular biology were misplaced, there might be large-scale projects that constrain the way biologists select hypotheses to develop in the first place. If such projects exist, then large-scale biology would be a threat to the independence of the 'small-lab' culture of molecular biology. These projects are 'centralizing' not in a spatial or institutional sense, but rather in an epistemic sense: they dictate the research agenda of small labs. One project that seems to fit this profile is TCGA (http://cancergenome.nih.gov/abouttcga/overview).

TCGA is a large-scale biomedical project organized as a consortium of several universities and hospitals. It was launched in 2005 by the NCI and NHGRI as a pilot project for a large-scale effort to map and characterize the molecular basis of tumors[15]. Like the HGP, it required a kind of 'regime' that is different from the one described by Knorr-Cetina in her ethnographic study of molecular biology, though not as centralized as some large-scale projects in physics. The consortium itself is organized around numerous centers, located geographically throughout USA (http://cancergenome.nih.gov/abouttcga/overview). The consortium serves many purposes and make use of different strategies. First, the general strategy of the consortium was to sequence genomes from thousands of cancer samples and organize the data into a systematic map of somatic mutations and structural variations (Ratti 2015). The underlying rationale for this is evolutionary. In principle, since driver mutations confer a growth advantage to cancer cells, we expect them to be positively selected. If this is the case, then selected mutations should be detected more often

---

[15] For a much more detailed history of TCGA, please see Plutynski's contribution to this volume

than passenger mutations. Therefore, the bigger the sample size of sequenced cancer samples grows, the more it is likely to detect mutations that are statistically significant[16]. It is important to emphasize the 'gene-hunting' nature of this enterprise. In accordance with the somatic mutation theory of cancer, the genome is the place to start to discover how cancer progresses. Scientists look for so-called 'cancer genes' (either oncogenes or tumor-suppressor genes), which work as 'fire starters' or dampeners of oncological processes at the molecular level. Because of the high-resolution analyses that can be performed, the emphasis in this paradigm of 'gene-hunting' is on mutations. Therefore, *TCGA aimed to uncover the genetic basis of cancer by sequencing thousands of samples and discovering driver mutations*.

The history of cancer genomics is intertwined with the history of TCGA. TCGA has corroborated the cancer genes that molecular oncology (typically a small science, as suggested by Weinberg (2014)) discovered over the last 30 years. However, due to the statistical power of its studies, it also uncovered processes involved in cancer development that had been ignored by small-based molecular oncology. A good example is a mutation in *IDH1* (Ledford 2010). This gene is involved in cell metabolism, which has been associated with cancer in the past (Pavlova and Thompson 2016), but was ignored by the molecular oncology tradition of the 1980s and 1990s. 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; smaller data sets would be unable to show a robust regularity. Scientists might be lucky and discover all the cancer genes with a piecemeal approach, but methodologically big numbers *do make a difference*.

This last point about luck and piecemeal approaches in choosing the right gene is important to understand the potential impact of a project such as TCGA. Finding the right gene to analyze is related to hypothesis generation; in typical decomposition/localization strategies in molecular oncology/biology, one begins with one or more genes of interest. These are experimentally manipulated to elaborate a mechanistic description of how those genes (when mutated) work either as oncogenes or tumor-suppressors. However, how does one choose the "right" gene? What kind

---

[16] Statistical significance ultimately depends on a cutoff chosen by investigators that would balance an expected rate of true-positives and false-negatives – with larger sample sizes providing a more favorable ratio (Park et al. 2010). Studies need to be larger to detect selected mutants that only exist in a subset of samples (Lawrence et al. 2014).

of evidence is needed to start an investigation about one gene rather than another? This is the problem of hypothesis generation. For current molecular biology with its connections to biomedicine, this is often the problem of identifying suitable *molecular targets* for drug development. 'Molecular target' refers to a broad set of biological entities (e.g. proteins, genes, SNPs) that are relevant to a disease, where relevance can be assessed in several ways. A leading review (Lindsay 2003) emphasized how target identification is geared towards a molecular approach: "an understanding of the cellular mechanisms underlying disease phenotypes of interest" (p 831). This aligns with the contribution of molecular biology to this enterprise and is a common assumption. For instance, a widely cited review on early phases of drug discovery associate target identification with basic research (in particular basic molecular research):

"[t]he initial research, often occurring in academia, generates data to develop a hypothesis that the inhibition or activation of protein or pathway will result in a therapeutic effect in a disease state. The outcome of this activity is the selection of a target which may require further validation prior to progression into the lead discovery phase in order to justify a drug discovery effort" (Hughes et al. 2011, p 1239).

Basic research in biomedicine – pursued mainly by small-scale molecular biology (*sensu* Weinberg or Alberts) – is aimed at identifying molecular targets. But how does one identify those targets? There are several possibilities. One can, again, just be lucky. Alternatively, one can work on a family of genes because his/her lab specializes on those genes. Biologists tend to set the starting point of their quest for targets or undertake 'gene-hunting' with *ad hoc* criteria (Patel et al 2013): "researchers have tended to work on a handful of favored genes, often identified in the literature by academic groups, amenable to low-throughput analysis" (Butcher 2003, p 367-368). In a recent study led by one of the author of the present chapter (Stoeger et al 2018), this observation has been systematically confirmed, though not from a historical point of view. In particular, the study observed that many genes are simply not mentioned in biomedical research. This could reflect unimportance or be related to "existing social structures of research, scientific and economic reward systems, medical and societal relevance, preceding discoveries, serendipity, the availability of technologies and reagents, and other intrinsic characteristics of genes" (p 2). Therefore, generally, hypotheses are generated (or, research questions are selected) by relatively opaque processes. This also reflects the epistemic independence of biological labs, in the sense

that each one has a tradition. Every laboratory leader has worked in previous labs, and he/she has created a lab by building on the tradition he/she knows from experience. This can reflect particular scientific questions, techniques, and biological entities that are the target of research[17].

Our hypothesis is that projects such as TCGA can potentially undermine the relative independence of small labs in the phase of hypothesis generation. Consider a fictional (but realistic) example (Ratti 2016). Imagine that we are looking for molecular targets (e.g., cancer genes, proteins, etc.) related to a chronic disease such as prostate cancer. There are at least three paths we can follow to identify them. First, we can start from genes that our lab has always worked on. This strategy would work if prostate cancer has always been the target of the lab; in a sense, we would be building on existing knowledge. Second, we could do an extensive literature search in PubMed. We will likely find interesting and detailed studies produced by small independent biological labs that, on the basis of cancer samples coming from two or three patients, identify targets of interest and construct detailed mechanisms on how these work. We might then choose targets that our lab is more familiar by drawing from studies that use methodologies similar to what we use. Third, we could look at TCGA database. Here we would learn of genes found to be mutated in all of the prostate cancer samples sequenced, their frequency, the kinds of mutations involved, and clinical information (inter alia). Most important, a platform such as TCGA would tell us *at the population level* which genes seem to be implicated in prostate cancer.

It is important to emphasize that TCGA is telling us two important things. First, TCGA is prescribing which research question we should ask. In fact, it is dictating which genes are more likely to be relevant for the kind of phenomenon we want to explain. But, more significantly, unlike the informal search on PubMed or the familiarity of our lab with a family of genes, TCGA claims that it has good reasons for dictating which gene we should choose. In order to think that a hypothesis about a molecular target is promising, we need evidence that such an entity has some role in as many individuals of that population as possible.

"when you know that in a specific subpopulation some genes have most of the time certain type of mutations, and that mutations are not present in a healthy population (i.e. the case of genome-wide

---

association studies), then you have good preliminary reasons to further investigate such mutations" (Ratti 2016)

By using TCGA, we are justified from the start in focusing on certain targets. Comprehensive publications from TCGA provide examples of efforts to discover cancer genes, validate them in both computational and experimental ways, and then provide a platform in which relevant information about these genes is publicly available for other labs. For example, Baley and colleagues (2018) analyze all TCGA exome data via 26 different bioinformatics tools. Next, outlier adjustments are performed and manual curation occurs where experts in the field scrutinize the biological plausibility of the genes and their properties using a variety of tools. Finally, an experimental validation is provided. Such an analysis would take months for a small lab with limited resources to perform, even for only a few genes. Given large-scale projects such as TCGA, small labs can just select one of those genes and start off their investigation of how exactly a gene is driver in a certain cancer. The large-scale projects such as TCGA provide *justificatory reasons* for what 'small labs' should undertake. If the reason to fund one small project instead of another depends on the solidity of the hypothesis proposed, then the robustness of statistical, computational, and experimental studies provided by projects such as TCGA seem eminently preferable to habits, intuitions and a literature search. Our hypothesis is that large-scale projects can be the main source of preliminary hypotheses that "small science" investigates.

4.1 Exploring the claim

TCGA has the potential of dictating which hypothesis or research question should be pursued: "*the viability of a project in small biology would be entirely dependent on the big biology infrastructure and its potential to bring into light promising target from an immense sea of data*" (Ratti 2016). To test this hypothesis, we designed a computational analysis to ascertain whether TCGA has *in fact* been the driving force described in the previous section. Given the cancer genes that we know today, we should observe a peak in their analyses starting in 2008, which is the publication year for TCGA pilot project (The Cancer Genome Atlas Research Network 2008). Subsequently, we expect to observe groups using TCGA resources to explore the nature of new cancer genes. It is difficult find informative proxies to test this claim. We contemplated two options.
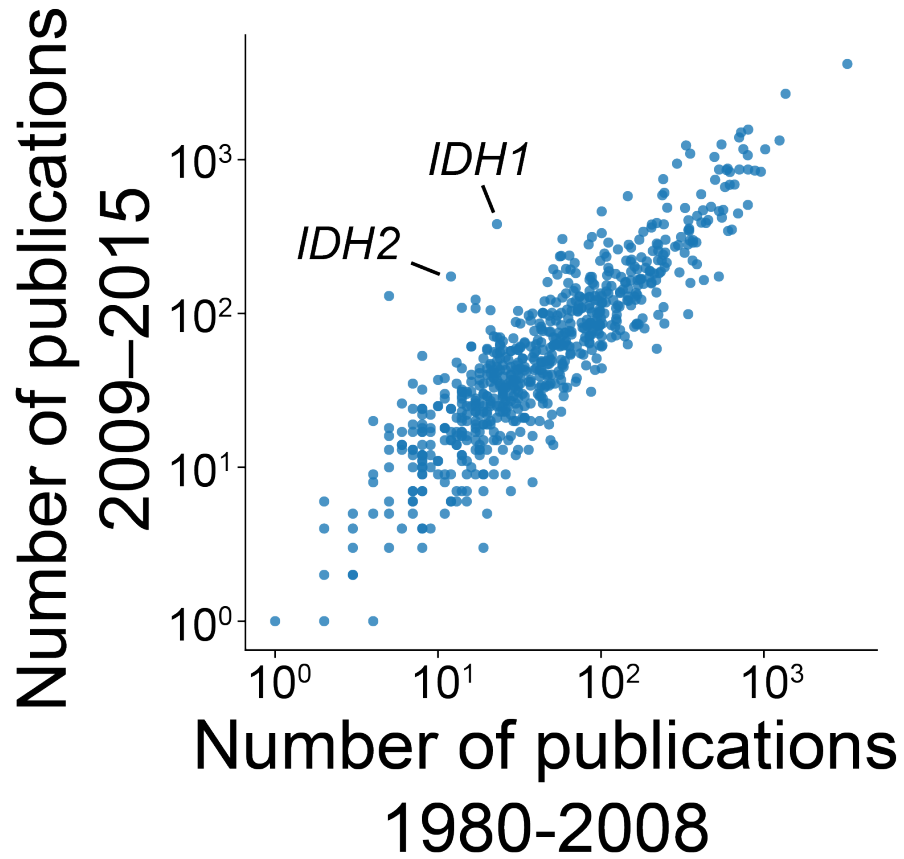
Figure 1: Occurence of cancer genes in Pubmed.
Number of publications in Pubmed containing cancer genes (blue dots) as defined by the Network of Cancer Genes project (Repana et al 2019). For visualization purposes we added +1 prior log10-transformation to show cancer genes without publications in the indicated timeframes. IDH1 and IDH2 are the genes *isocitrate dehydrogenase 1* and *isocitrate dehydrogenase 2*.

First, we could analyze when known cancer genes were cited. If TCGA affected the prioritization of cancer genes for small laboratories, some cancer genes should disproportionately appear in the biomedical literature after 2008. Using this approach, we obtained the mapping between publications on PubMed and individual human protein-coding genes from the National Center for Biotechnology Information, and added the publication year as listed in Medline. In parallel, we obtained a list of currently known cancer genes from the *Network of Cancer Genes* project (Repana et al 2019). Next, we counted the number of occurrences in publications for each cancer gene within the publications published until the year 2008 (the year of the first publication

of TCGA) and between the years 2009 and 2015 (the last year for which bibliometric data is available). Comparing the number of publications in these two time-frames, we observed a strong correlation (Figure 1, Spearman: 0.88), which suggested that the extent of scholarship on individual cancer genes had changed little after the first publication of TCGA. Notable exceptions to the general trend are *IDH1* and *IDH2*. The role of mutated *IDH1* in multiple cancer types was described and popularized between 2006 and 2010 by a series of publications (Sjoblom et al. 2006; Zhao et al 2009; Paschka et al 2010; Yan et al 2009), which included one publication of the CGA (Verhaak et al 2010). This finding supports our hypothesis that large-scale efforts can dictate subsequent small-scale research on individual genes, *but also demonstrates that these are rare events*.
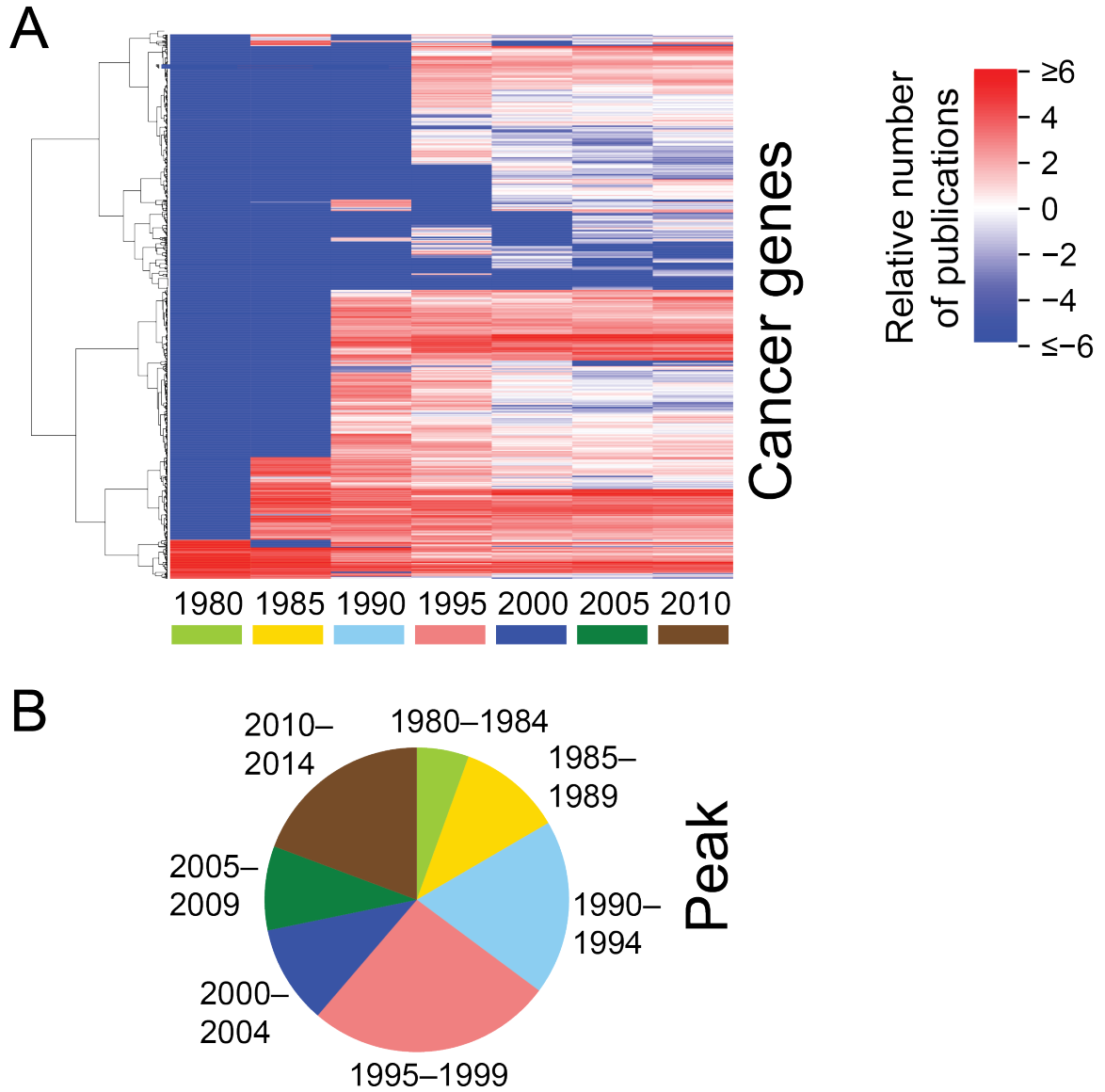
Figure 2: Peak times of contribution of cancer genes toward annual literature on human protein-coding genes. (A) Clustermap of contribution of individual cancer genes toward annual human literature on Pubmed. Years indicate the starting year of a 5-year interval. To account for changes in the total number of publications on Pubmed, we normalized the number of publications relative to the number expected if each protein-coding gene would be studied equally. (B) Fraction of cancer genes, which had their highest relative contribution to Pubmed within individual time intervals of (A).

We extended our analysis backward in time to see if scholarship on cancer genes might have been more dynamic in the past by computing the relative contribution of each cancer gene to the literature on human genes in each year. To avoid counting single publications with several genes multiple times, we counted publications containing multiple genes as 1/(number of contained genes) toward the count for each of the contained genes. Finally, we obtained for each 5-year observational period the median contribution of each gene, and normalized it with a naive null model in which each human protein-coding gene would be studied equally. We did the normalization by taking the ratio of the observed annual contribution and the contribution anticipated by the naive null model, and subsequently performed a log2-ratio. On this analysis, a value of +1 would indicate that a given cancer gene would be studied twice as much compared to a culture where researchers would randomly publish on individual genes, whereas a value of -1 would indicate that a given cancer gene would be studied half as much. Recapitulating our earlier observation, we see that different cancer genes tend to be studied to a varying extent (Figure 2). However, we also observe that 86% of the cancer genes have a log2 ratio above 0 in at least one time interval, and 60% do so for the most recent time interval. This suggests that, cumulatively, small laboratories study cancer genes more frequently than other genes.

Detailing the temporal changes, we see that *only 19% of the cancer genes peaked after 2010*. Instead, we notice that *61% of all cancer genes peaked before the year 2000*, with the time span between 1995 and 1999 having been the peak time for the largest amount of cancer genes (26%). This demonstrates that the *most recent years, which coincided with TCGA, did not lead to a surge of research on cancer genes*. Another conclusion that we can draw from this analysis is that most cancer genes were already known prior the completion of the HGP and the start of TCGA, with several known in the 1980s continuing to be among the most studied cancer genes more recently. We suspect that the decreasing contribution of these genes toward the annual literature may be explained by these cancer genes having been identified early on, while only a small subset of human genes had been identified yet; subsequently, they contributed less to the annual literature as new cancer and non-cancer genes were being identified and studied.
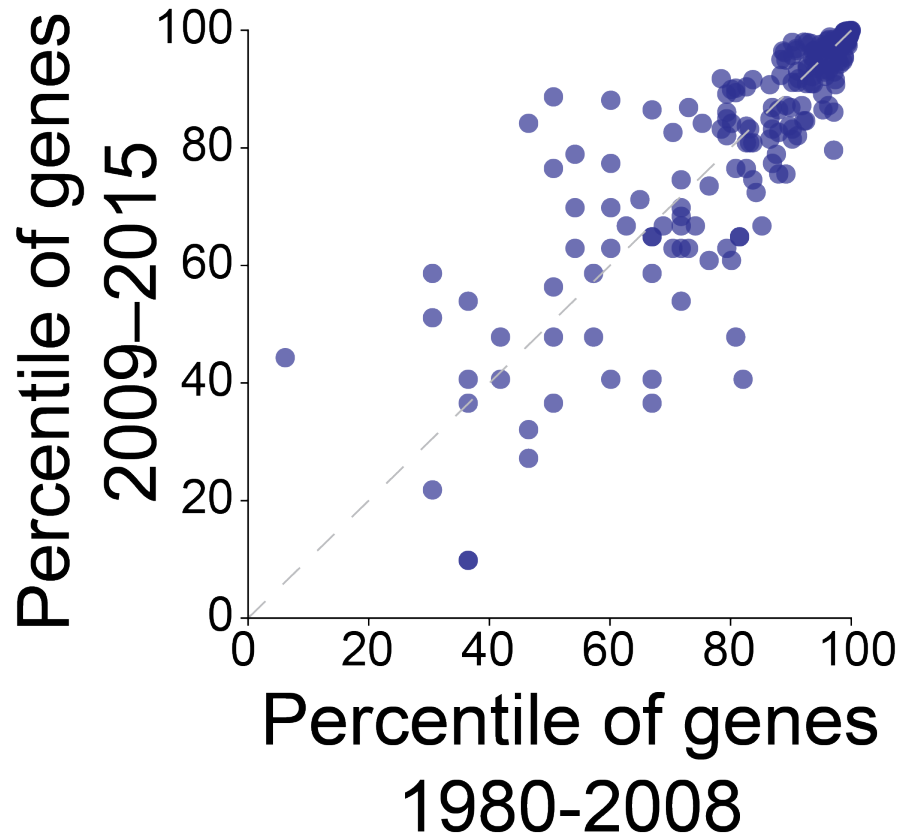
Figure 3: Absent rise in literature for genes reported by TCGA in 2008.
We counted the number of publications for each human protein-coding gene, and for each time period independently ranked all genes according to the number of publications. Shown are percentiles of these ranks for all cancer genes reported as verified in TCGA 2008 (Cancer Genome Atlas Research Network 2008) Supplemental Figure 6 (purple). The observed differences are not significant (Wilcoxon: 0.49)

The second option for an illuminating proxy to probe the impact of TCGA on research involved restricting ourselves to 223 genes that had been reported and experimentally verified within the first 2008 publication of results by TCGA. These were then tested to see whether they would become more or less studied over time when compared to all other genes (Figure 3). Using the same time frames for the number of publications per year, we compared cancer genes against all other genes by ranking genes according to the number of publications; the gene at the 100 percentile corresponds to the gene occurring in the most publications. For both groups, the median percentile of all 223 cancer genes is 92%, confirming that cancer genes are already heavily studied when compared to all other genes. However, we did not observe any difference in the ranks when

comparing the two time-frames (Wilcoxon: 0.49). This too suggests that TCGA had a surprisingly small impact on the genes studies in cancer research.

4.2 Why TCGA has not been an epistemically centralizing effort

There may be several explanations for why we did not confirm our initial hypothesis. First, a concern could be raised that we did not consider the technological advancements brought forth by projects such as TCGA, and this may be the most important contribution of the project. As Eddy (2013) claimed, there are no 'big experiments' in biology; large-scale projects yield both maps and ways of stimulating technological development in a field. In this respect, TCGA's most important contribution lies in the way it has developed new tools and technology. This is a constant in projects such as HGP or TCGA – "[g]enome research thus generated not only scientific products, but also the means for their production" (Mueller-Wille and Rehinberger 2012, p 201). Plutynski (this volume) notices that "there were in fact many different kinds of goals at play in TCGA", and in addition to scientific goals, "TCGA promoted improvements in speed, accuracy, and lower cost of sequencing" (p 5). Improving sequencing technologies is just another way in which large-scale projects can be 'facilitators' by improving the hypothesis development phase for small labs.

Another reason is more empirical. Since the majority of cancer genes were discovered between 1980 and 1999 (and hence before the cancer genomics era), it just turns out that the number of cancer genes is limited and researchers were lucky to discover many of them before the advent of next-generation sequencing. Therefore, we should not expect a sudden increase in the number of cancer genes studied after 2008 because most had already been discovered. According to this hypothesis TCGA guided small-scale studies, but only for such a small set of cancer genes that it would have been missed by our statistical analysis – or by enabling conceptually distinctive hypotheses that transcend from individual genes. Moreover, we cannot exclude the possibility that overall research on cancer genes might have decreased without TCGA due to competition with other fields of biomedical research. Also, our hypothesis might have been incomplete in regards to TCGA because the disease-relevance of genes is not by itself sufficient to dictate subsequent small-scale studies[18].

---

[18] A further possibility could be that TCGA shifted the formation of hypothesis away from individual cancer genes toward hypothesis that are formed around systems of cancer genes (Marcum et al. 2008). Due to space constraints, we excluded a more detailed discussion

But our initial hypothesis is not without merit, and it may turn out to be more accurate in the future. As mentioned in the introduction, there are on-going attempts to undermine small-labs autonomy in the way we described. One example is the Illuminating the Druggable Genome Consortium (IDG). This consortium promotes research on rarely studied genes that belong to gene families that are considered to be important and malleable to pharmacological interference. IDG is supported by the Common Fund Program of the NIH and is formed by multiple academic centers and collaborates with the National Center for Advancing Translational Science. Together they follow several strategies. First, they provide information resources (Nguyen et al. 2017). Second, they solicit funding calls for specific genes prioritized by them[19]. Third, they engage in community outreach and partner with Nature Reviews Drug Discovery to regularly highlight the genes prioritized by IDG to their readers (https://commonfund.nih.gov/IDG/NRDD). Another, not yet formalized, example stems from the International Mouse Phenotyping Consortium, a large-scale consortium that mutates mouse genes to monitor mice for phenotypic changes. One of their insights is that many genes that affect the physiology of mice have not been subjected to small-scale studies. Besides publicizing the genes on social media (https://twitter.com/impc), this consortium, and like-minded biomedical researchers, also recently called for a deep-genome project to characterize all human genes and their mouse orthologs in detail (Lloyd et al. 2020). Our analysis framed in the tripartite account of discovery shows a way in which it is possible to monitor how large-scale biology is impacting small-scale biology.

CONCLUSION

In this chapter, we have discussed the relation between large-scale projects and 'small labs' within the knowledge culture of experimental life sciences, with a focus on molecular biology and genomics. In particular, we have discussed ways in which a large-scale project can influence the discovery strategies of small biological labs. We have articulated two ideas.

First, large-scale biology can shape discovery strategies by contributing to the ways biologists develop their initial hypotheses. We named these large-scale projects *facilitators*. In particular, projects such as HGP and TCGA can provide means and tools (sequences, computational infrastructures such as databases, faster and cheaper sequencing technologies, etc)

---

[19] See for instance, https://grants.nih.gov/grants/guide/rfa-files/rfa-rm-18-021.html, https://grants.nih.gov/grants/guide/rfa-files/RFA-RM-19-012.html

to aid biologists in developing initial conjectures or research questions. Facilitators help biologists to get where they want to get faster and more efficiently.

But we have also hypothesized that large-scale projects can do something more radical. In particular, large-scale projects can dictate the research agenda of small labs, by shaping the phase of hypothesis generation. In this sense, large-scale project can be epistemically centralizing, because they 'centralize' how hypotheses are formulated in the first place. Given the particular epistemic structure of molecular oncology and cancer genomics, we conjectured that TCGA may be an epistemically centralizing project. By providing high-resolution maps of whole-genome of cancer samples, TCGA is in a better position to identify candidates cancer genes in a much more systematic way. If this is true, then TCGA can dictate and determine which are the best biological hypotheses to pursue further with the typical means of molecular biology small labs. However, an analysis of molecular oncology and cancer genomics literature from 1980 have provided evidence that this did not happen. Therefore, our hypothesis that TCGA can function as an epistemically centralizing project does not stand up to careful (empirical) scrutiny. But we have also noted, at the very end, that newer large-scale projects have been designed explicitly to become epistemically centralizing.

REFERENCES

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

Barnes, B., & Dupré, J. (2008). *Genomes and What to Make of Them*. Chicago & London: University of Chicago Press.

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

Boem, F., & Ratti, E. (2016). Towards a Notion of Intervention in Big-Data Biology and Molecular Medicine. In G. Boniolo & M. Nathan (Eds.), *Philosophy of Molecular Medicine: Foundational Issues in Research and Practice* (pp. 147–164). London: Routledge.

Butcher, S. P. (2003). Target Discovery and Validation in the Post-Genomic Era. *Neurochem Res*, *28*(2), 367–371.

Conesa, Ana, and Ali Mortazavi. 2014. "The Common Ground of Genomics and Systems Biology." *BMC Systems Biology* 8 (Suppl 2): S1. doi:10.1186/1752-0509-8-S2-S1.

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.

Duhem, P. (1954). *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*.

Eddy, S. R. (2013). The ENCODE project: missteps overshadowing a success. *Current Biology : CB*, *23*(7), R259–61. http://doi.org/10.1016/j.cub.2013.03.023

Franklin, L. R. (2005). Exploratory Experiments. *Philosophy of Science*, *72*(December), 888–899.

Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford University Press.

Hacking, I. (1983). *Representing and Intervening - Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press.

Hardwig, J. (1985). Epistemic Dependence. *The Journal of Philosophy*, *82*(7), 335–349.

Hilgartner, S. (2013). Constituting large-scale biology: Building a regime of governance in the early years of the Human Genome Project. *BioSocieties*, *8*, 397–416. http://doi.org/10.1057/biosoc.2013.31

Hilgartner, S. (2017). *Reordering Life*. MIT Press.

Hughes, J. P., Rees, S. S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British Journal of Pharmacology*, *162*(6), 1239–1249. http://doi.org/10.1111/j.1476-5381.2010.01127.x

Kevles, D. (1997). Big Science and Big Politics in the United States : Reflections on the Death of the SSC and the Life of the Human Genome Project. *Historical Studies in the Physical and Biological Sciences*, *27*(2), 269–297.

Knorr-Cetina, K. (n.d.). *Epistemic Cultures*. Cambridge, MA: Harvard University Press.

Lawrence, Michael S., Petar Stojanov, Craig H. Mermel, James T. Robinson, Levi A. Garraway, Todd R. Golub, Matthew Meyerson, Stacey B. Gabriel, Eric S. Lander, and Gad Getz. 2014. "Discovery and Saturation Analysis of Cancer Genes across 21 Tumour Types." *Nature* 505 (7484). Nature Publishing Group: 495–501. doi:10.1038/nature12912.

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

Leng, Gareth, and Rhodri Ivor Leng. 2020. *No TitleThe Matter of Facts: Skepticism, Persuasion, and Evidence in Science*. The MIT Press.

Lindsay, M. a. (2003). Target discovery. *Nature Reviews. Drug Discovery*, *2*(10), 831–8. http://doi.org/10.1038/nrd1202

Lloyd, K. C.Kent, David J. Adams, Gareth Baynam, Arthur L. Beaudet, Fatima Bosch, Kym M. Boycott, Robert E. Braun, et al. 2020. "The Deep Genome Project." *Genome Biology* 21 (1): 1–6. doi:10.1186/s13059-020-1931-9.

Morange, Michel. 2020. *The Black Box of Biology: A History of the Molecular Revolution*. Harvard University Press.

Nguyen, Dac Trung, Stephen Mathias, Cristian Bologa, Soren Brunak, Nicolas Fernandez, Anna Gaulton, Anne Hersey, et al. 2017. "Pharos: Collating Protein Information to Shed Light on the Druggable Genome." *Nucleic Acids Research* 45 (D1): D995–1002. doi:10.1093/nar/gkw1072.

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.

Park, Ju Hyun, Sholom Wacholder, Mitchell H. Gail, Ulrike Peters, Kevin B. Jacobs, Stephen J. Chanock, and Nilanjan Chatterjee. 2010. "Estimation of Effect Size Distribution from Genome-Wide Association Studies and Implications for Future Discoveries." *Nature Genetics* 42 (7): 570–75. doi:10.1038/ng.610.

Paschka, Peter, Richard F. Schlenk, Verena I. Gaidzik, Marianne Habdank, Jan Krönke, Lars Bullinger, Daniela Späth, et al. 2010. "IDH1 and IDH2 Mutations Are Frequent Genetic Alterations in Acute Myeloid Leukemia and Confer Adverse Prognosis in Cytogenetically Normal Acute Myeloid Leukemia with NPM1 Mutation without FLT3 Internal Tandem Duplication." *Journal of Clinical Oncology* 28 (22): 3636–43. doi:10.1200/JCO.2010.28.3762.

Patel, M. N., Halling-Brown, M. D., Tym, J. E., Workman, P., & Al-Lazikani, B. (2013). Objective assessment of cancer genes for drug discovery. *Nature Reviews Drug Discovery*, *12*(1), 35–50. http://doi.org/10.1038/nrd3913

Pavlova, N. N., & Thompson, C. B. (2016). The Emerging Hallmarks of Cancer Metabolism. *Cell Metabolism*, *23*(1), 27–47. http://doi.org/10.1016/j.cmet.2015.12.006

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

Ratti, E. (2016). The end of "small biology"? Some thoughts about biomedicine and big science. *Big Data & Society*.

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

Repana, Dimitra, Joel Nulsen, Lisa Dressler, Michele Bortolomeazzi, Santhilata Kuppili Venkata, Aikaterini Tourna, Anna Yakovleva, Tommaso Palmieri, and Francesca D. Ciccarelli. 2019. "The Network of Cancer Genes (NCG): A Comprehensive Catalogue of Known and Candidate Cancer Genes from Cancer Sequencing Screens." *Genome Biology* 20 (1). Genome Biology: 1–12. doi:10.1186/s13059-018-1612-0.

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

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

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

Richardson, S., & Stevens, H. (Eds.). (2015). *Postgenomics - Perspective on Biology After the Gnome*. Duke University Press.

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

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

Stevens, H. (2013). *Life out of sequence - A data-driven history of bioinformatics*. Chicago: Chicago University Press.

Stoeger, T., Gerlach, M., Morimoto, R. I., & Nunes Amaral, L. A. (2018). Large-scale investigation of the reasons why potentially important genes are ignored. *PLOS Biology*, *16*(9), e2006643. http://doi.org/10.1371/journal.pbio.2006643

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*, *40*(15), 1–12. http://doi.org/10.1093/nar/gks596

Verhaak, Roel G.W., Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, C. Ryan Miller, et al. 2010. "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1." *Cancer Cell* 17 (1). Elsevier Ltd: 98–110. doi:10.1016/j.ccr.2009.12.020.

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

Waters, Kenneth. 2008. "Beyond Theoretical Reduction and Layer-Cake Antireduction: How DNA Retooled Genetics and Transformed Biological Practice." In *The Oxford Handbook of Philosophy of Biology*, edited by Michael Ruse. Oxford University Press.

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

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

Westerhoff, Hans, and Bernhard Palsson. 2004. "The Evolution of Molecular Biology into Systems Biology." *Nature Biotechnology* 22: 1249–52.

Yan, Hai, D. Williams Parsons, Genglin Jin, Roger McLendon, B. Ahmed Rasheed, Weishi Yuan, Ivan Kos, et al. 2009. "Mutations in Gliomas." *New England Journal of Medicine* 360 (8): 765–73. doi:10.1056/NEJMoa0808710.Zhao, S, Y Li, and W Xu. 2009. "Glioma-Derived Mutations in IDH1 Dominantly Inhibit IDH1 Catalytic Activity and Induce HIF-1a." *Science* 284 (15): 9835–44. doi:10.1074/jbc.M807084200.