Big Data Biology: Between Eliminative Inferences and Exploratory Experiments

Emanuele Ratti*†

Recently, biologists have argued that data-driven biology fosters a new scientific methodology, namely, one that is irreducible to traditional methodologies of molecular biology defined as the discovery strategies elucidated by mechanistic philosophy. Here I show how data-driven studies can be included in the traditional mechanistic approach in two respects. On the one hand, some studies provide eliminative inferential procedures to prioritize and develop mechanistic hypotheses. On the other, different studies play an exploratory role in providing useful generalizations to complement the procedure of prioritization. Overall, this article aims to shed light on the structure of contemporary research in molecular biology.

1. Introduction. In recent years, a debate has emerged among biologists concerning how sequencing technologies are influencing and changing the methodologies of molecular biology. Along with other high-throughput methods, sequencing technologies have been increasingly tied to the so-called big data issue. According to some commentators in the field, big data are "fostering a new mode of scientific research, which some commentators refer to as 'data-driven'" (Leonelli 2012b, 1). Moreover, some argue that data-driven approaches in molecular biology are going to replace more 'traditional' methodologies. This claim has been the focus of considerable controversy in the past few years (Brenner 1999; Golub 2010; Weinberg 2010;

Received June 2014; revised November 2014.

*To contact the author, please write to: Dipartimento di Scienze della Salute, Universita' di Milano, Department of Experimental Oncology, Istituto Europeo di Oncologia, Milan, Italy; e-mail: emanuele.ratti@ieo.eu.

†I am extremely grateful to the fellows of the PhD program FOLSATEC (especially Pierre Luc Germain, Federico Boem, Marco Annoni, and Giovanni Boniolo) for their valuable comments on previous drafts, as well as to Michael Weisberg and his group in Philadelphia (especially Emily Parke, Carlos Santana, Jane Reznik, and Alkistis Elliot-Graves). I am in debt also to David Teira for his indispensable comments on advanced drafts.

Philosophy of Science, 82 (April 2015) pp. 198–218. 0031-8248/2015/8202-0003$10.00

Alberts 2012; Garraway and Lander 2013; Vogelstein et al. 2013). 'Data-driven' research is understood as a 'hypothesis-free' methodology,[1] while 'traditional' molecular biology methodologies are taken to be 'hypothesis-driven' (Brenner 1999; Weinberg 2010; Alberts 2012). As emphasized by the received view in philosophy of biology, traditional methodologies of molecular biology can be understood in terms of the discovery strategies illustrated by the so-called mechanistic philosophy (Craver and Darden 2013).

This controversy is of philosophical interest for a number of reasons. Here I am interested in one in particular. As emphasized above, the perspective of 'mechanistic philosophy' constitutes most of the methodological core of molecular biology. Since data-driven molecular biologists argue that their methodologies are somehow irreducible to traditional approaches, there is an obvious question to ask: is data-driven molecular biology a new approach to discovery in biology that mechanistic philosophy cannot account for?

I argue that the answer to this question is negative by developing a proposal informally put forth in the past few years. Recently, some philosophers and historians have proposed that there is not a dichotomy between data-driven and hypothesis-driven, but rather the approaches are 'hybridized' (Smalheiser 2002; Kell and Oliver 2003; Strasser 2011; Keating and Cambrosio 2012; Leonelli 2012a; O'Malley and Soyer 2012). The core of the 'hybridization' proposal rests on two tenets. First, data-driven strategies aim to generate hypotheses. Next, hypotheses are developed and tested with procedures that are consistent with the epistemic program of 'mechanistic philosophy'. Therefore, the hybridization of data-driven and hypothesis-driven is compatible with the mechanistic program. I will use the acronym 'DDHD' to refer to the hybridization of data-driven and hypothesis-driven research. However, DDHD faces two problems. First, the proposal is mostly informal. Therefore, it is not entirely clear how DDHD is compatible with more traditional approaches to molecular discovery. Second, the proposal includes only a part of studies that are labeled as data-driven. In particular, the proposal encompasses only those screenings that make use of sequencing technologies in order to identify entities that might possibly play a causal role in a phenomenon. However, there are data-driven studies (which I call "mining studies") characterized by the analysis of data generated by big scientific projects such as the Encyclopedia Of DNA Elements (ENCODE) project that do not bear any straightforward similarity to DDHD. Therefore, it is controversial whether mining studies can be subsumed into a more traditional perspective.

Accordingly, the aim of this article is twofold. First, I show how DDHD is compatible with the broader approach of mechanistic philosophy. Next, I argue that mining studies play an important role in DDHD. In particular,

1. As it has been rhetorically stated (Golub 2010), data should come first.

https://doi.org/10.1086/680332 Published online by Cambridge University Press

mining studies play the role of exploratory experiments in the discovery strategies of DDHD. Therefore, the overall goal of this paper is to grasp the structure of contemporary research in molecular biology.

The structure of the paper is as follows. Section 2 deals with two key examples of data-driven science that can be easily reduced to the DDHD proposal. Next, I explain how in these studies DDHD and mechanistic philosophy are compatible, highlighting the structure of their integration. In particular, in sections 2.1 and 2.2 I reconstruct the structure of data analysis and hypothesis development. In section 2.3 I show how these data-driven studies are perfectly compatible with hypothesis-driven approaches. In section 3 I turn to mining studies and elucidate their role in DDHD.

2. Data-Driven Research and Hypothesis-Driven Research. In order to disentangle the controversy presented in the introduction, in this section I reconstruct the structure of DDHD and show how this approach is compatible with the epistemic perspective of mechanistic philosophy. I should note that this is an interpretation of the practices of DDHD. In particular, I emphasize that DDHD is constituted by three phases:

1. formulation of an initial set of competing hypotheses;
2. elimination of false (or less probable) hypotheses;
3. test (validation) of hypotheses not eliminated in phase 2.

Phases 1 and 2 bear resemblances to the 'eliminative inductive' framework (Earman 1992; Kitcher 1993; Norton 1995). Eliminative induction is also known as 'induction by means of deduction' (Hawthorne 1993), 'eliminative inference' (Forber 2011), or 'strong inference' (Platt 1964). Among the alternatives, I prefer to use the more neutral label 'eliminative inference' (Forber 2011) because it is not clear exactly which inductive characteristics this process involves. In eliminative inferences a set of premises stimulates a (finite) universe of competing hypotheses. Premises might be characterized as a 'prior state of individual practice' (Kitcher 1993) that practitioners use to select candidate theories/hypotheses. Next, with the help of other premises and new evidence, hypotheses are progressively discarded, until only one remains: the true hypothesis. Forber (2011) correctly points out that eliminative inferences are traditionally characterized as a method of theory choice, in the sense that the process of elimination determines which hypothesis is true and which are false. However, in scientific practice evidence seldom provides a strict deductive elimination of hypotheses. Rather, evidence provides statistical support. As Forber states, "Perhaps eliminative inferences do not make theory choices but establish the boundaries for such choices" (2011, 192). Therefore, I think it is better to frame eliminative inferences as a process of prioritization of theories and hypotheses rather than theory choice.

Moreover, 'prioritization' implies the idea of an additional procedure aimed at providing more evidence for what has been prioritized. This is exactly the process that occurs in phase 3, that is, the phase where scientists look for more stringent evidence for prioritized hypotheses. Therefore, one might frame DDHD phases as follows:

1. the generation of a preliminary set of hypotheses from an established set of premises;
2. the prioritization of some hypotheses and discarding of others by means of other premises and new evidence;
3. the search for more stringent evidence for prioritized hypotheses.

DDHD is mainly a process of prioritization, and it fits the account of eliminative inferences.

In DDHD, 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 2 should be validated in phase 3, in the sense that the way entities causally contribute to the phenomenon of interest should be clearly identified. The causal role of entities in the phenomenon of interest is framed in terms of the contribution of entities to biological mechanisms that produce the phenomenon of interest.

Finally, it is worth emphasizing that the structure I have identified does not capture the structure of all big data studies. The hybridization proposal does not encompass traditional big data studies in biology, such as the Human Genome Project, or the mining studies that I discuss in section 3. The account of DDHD I am developing is only meant to apply to a specific class of studies exemplified in this article by genome-wide association studies (GWASs) and cancer genomics.

Cancer genomics is a molecular approach to cancer. The advantage over traditional molecular studies of cancer is that cancer genomics approaches the discovery of cancer genes not in a sparse fashion, but rather by generating in the first instance a systematic view of all the mutated genes (and all the mutations) of a cancer genome.

In contrast, GWASs fall in the category of genetics epidemiology. They aim at scanning markers across the entire genome of many individuals in order to find variants associated with common diseases.

The reason for choosing GWASs and cancer genomics is that these two endeavors have only become possible after the establishment of post–Human Genome Project technologies, and, consequently, they share several of the features usually ascribed to big data. Both generate vast data sets, and both are supposed to exemplify a 'hypothesis-free' way of doing science (Guessous, Gwinn, and Khoury 2009; Brookfield 2010; Garraway and Lander 2013; Vogelstein et al. 2013). Once the structure of these two types of screenings is clarified, I shall compare it to the kind of approach endorsed by 'traditional' molecular biologists.

2.1. Establishing the Initial Universe of Hypotheses. In this subsection I explain in detail phase 1 in DDHD. This is the phase in which a set of premises is used to draw the boundaries of a 'universe' of competing hypotheses. Hypotheses are about the causal contribution (to phenomena) of entities.

The first point to note is that phase 1 is not a hypothesis-free step. The impossibility of completely 'hypothesis-free' scientific research is vastly acknowledged in the philosophy of science literature (Rheinberger 2011; Leonelli 2012a). The general idea is that data cannot be gathered without the guidance of antecedent hypotheses because one would have no basis on which to identify relevant data without at least preliminary guidelines. Therefore, data-driven research must make use of hypotheses of various kinds. I call these hypotheses "background assumptions."

Background assumptions in data-driven studies play a weaker role than hypotheses or theories might play in the standard hypothetico-deductive method. To clarify this point, I use the distinction between theory-driven and theory-informed (Waters 2007). On the one hand, an experiment is theory/hypothesis-driven when theories/hypotheses influence directly the experimental design in order to answer a specific question, like the test of theories or hypotheses themselves. On the other hand, an experiment is theory-informed when theories or hypotheses do not provide specific expectations or anticipations on the results that will be discovered, and experimental designs are not set up in order to generate a specific effect. 'Theory-informed' experiments are used not to test a preexisting theory but rather to provide guidelines and suggest strategies in order to foster the discovery of significant findings about a phenomenon when a predefined theory is absent. Phase 1 in DDHD is theory-informed but not theory-driven. By providing guidelines, background assumptions establish which data are relevant and which are not. By doing this, background assumptions also specify the kind of hypotheses that will compose the preliminary universe of hypotheses.

There are two kinds of background assumptions. The first set of assumptions concerns the particular scientific problem stimulating a research. Without a scientific problem, no research would take place because nothing would

motivate the research. Therefore, a scientific problem "informs" a team of scientists in the sense that it provides at least a rough idea of the direction that research should take. Next, there is a second loose guideline, that is, a tentative solution to the problem. Some classes of problems can be studied by a number of different disciplines simultaneously. Take, for instance, the study of perception and cognition. Peschard and van Fraassen (2014) notice that computational approaches differ from robotic programs in the choice of the metaphor to characterize the cognitive system. For instance, the computational program grounds its research on the metaphor that the cognitive system is a computer, while the robotic approach grounds it on the metaphor of cognitive systems as embodied and embedded in the environment. These metaphors encompass the kinds of background assumptions I am talking about. They are composed of the same scientific problem ('how does the cognitive system work?') and different tentative answers to it ('it is a computer/it is embodied'). Tentative answers trace the boundaries where modeling strategies move.

Background assumptions of DDHD have a peculiar feature. They not only suggest relevant data but also draw the boundaries of an initial set of abstract hypotheses.

GWAS practitioners frame their research exactly in light of a scientific problem and tentative answers (Kitsios and Zintzaras 2009). The scientific problem can be easily identified: GWASs aim to discover the genetic basis of common diseases. However, this is the general aim of genetics epidemiology. GWASs are peculiar because of the tentative answers provided to solve this problem. First, genetic variations are proxies for the genetic basis of common diseases. This assumption is complemented with another concept: one should be interested in single nucleotide polymorphisms (SNPs) and not in other types of variants. An SNP is a single nucleotide variant whose allele is present in at least 1% of the population. In other words, a scientific question and tentative answers prescribe facts of interest. These assumptions inform GWASs, in the sense that they supply guidelines to select relevant data. Moreover, by identifying relevant data from background assumptions, one can also draw the boundaries of a (finite) universe of competing hypotheses: all the SNPs initially detected by an SNP array might be, potentially, causal variants. Since an SNP array may detect up to two million SNPs, the initial universe of hypotheses of a GWAS might be composed by millions of very abstract hypotheses like '*x* might have a causal role in the development of *y*'.

Cancer genomics also has a number of background assumptions. The scientific problem in question is to understand how tumors develop. The first tentative answer to this problem 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 generally interested in somatic (ac-

quired) mutations.[2] Moreover, cancer genomics is interested only in driver mutations (mutations that drive cancer development in the first place, by providing selective advantages to the cells that carry them), rather than so-called passenger mutations (bystander mutations not influencing cancer development). Therefore, cancer genomics will look for driver mutations within a set of specific facts, that is, somatic mutations. The initial (somatic) mutations detected constitute the universe of hypotheses that will be narrowed by eliminative inferences.

2.2. Eliminating Hypotheses. In phase 2 'eliminative principles' are used to narrow the finite universe of hypotheses, that is, to eliminate false (or less probable) hypotheses.

As emphasized in section 2.1, by means of background assumptions GWAS practitioners establish an initial universe of hypothesized entities that may be responsible for a particular phenotype. However, any epidemiologist knows that most of the SNPs cannot be responsible for the phenotype. Therefore, epidemiologists need some sort of criteria to identify SNPs that are not causal variants. There are two types of criteria. The first is based on a statistical analysis. In a typical GWAS a group of individuals with the phenotype of interest is compared with 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 (named 'significance level'[3]). To put it in simple terms, if the proportion between the frequencies in the two groups of a particular allele of an SNP exceeds the significance level in favor of the group with the phenotype of interest, then the variation is taken to be associated with the disease. If an SNP has the allele frequency below the significance level, then it is discarded as spuriously associated with a disease.[4] Other statistical procedures called 'technical derivation' and 'replication' (Hunter, Altshuler, and Rader 2008) might be used in order to refine the main statistical analyses.

2. This is an assumption of certain 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).

3. How to choose the significance level is a matter of debate, and it varies according to the particular experimental design employed, sample size, etc.

4. It is important to stress that SNPs are not discarded tout court. An SNP discarded in a study might not be eliminated in studies with bigger sample sizes for technical statistical reasons.

https://doi.org/10.1086/680332 Published online by Cambridge University Press

The second criterion is employed not only to eliminate other hypotheses but also to develop more probable guesses. This criterion is biologically driven (Boyle et al. 2012; Schaub et al. 2012). 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 fall in noncoding regions. The identification of functions of noncoding regions is a challenging endeavor. To deal with this issue, the ENCODE project (ENCODE Project Consortium 2012) has provided annotations for all the biochemical activities within the human genome at a nucleotide resolution (Germain, Ratti, and Boem 2014). This means that it is possible to see whether SNPs that were not eliminated (read: that are prioritized) in the previous phases locate in noncoding regions that "overlap a functional region or are in strong linkage disequilibrium with a SNP overlapping a functional region" (Schaub et al. 2012, 1749). If an SNP falls in a region of the genome and the biochemical function of this locus has nothing to do with the phenotype investigated, then the SNP can be eliminated from the universe of hypotheses. If an SNP occupies a region that is annotated as, for example, a transcription factor binding site, then scientists might make the hypothesis that the SNP is actually regulating a gene. 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, 819).

With this procedure, variants that are spuriously associated with the phenotype of interest are eliminated, while prioritized SNPs correspond to the final universe of entities hypothesized to be responsible for the phenotype of interest.

Similar procedures may be drawn for cancer genomics as well (Raphael et al. 2014). Statistical analysis is the first procedure. A common view in the literature is that, 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, 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" means. This is why, in each high-throughput screening, statisticians calculate a background mutation rate (BMR). The idea is that a mutation, in order to be a candidate driver mutation, should be present at a rate that is higher than the BMR. If this is not the case, 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 the BMR.[5] The

5. As for GWASs, in cancer genomics a mutation might be discarded in a study, but its frequency might be above the BMR in studies with a bigger sample.

statistical analysis on mutations is complemented with a statistical analysis on genes. Candidate driver mutations are likely to target genes that are mutated at a higher rate than a BMR designed specifically for genes.

Next, there is a biologically driven procedure. In order to eliminate mutated genes (and, as a consequence, other mutations), a typical standard is to check whether recurrently mutated genes overlap with known cancer pathways (Vandin, Upfal, and Raphael 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 et al. (2013) eliminate several recurrent mutations (by eliminating recurrently mutated genes) from the universe of initial hypotheses because they do not participate in any known cancer pathways.[6]

2.3. The Phase of Hypothesis Testing and Final Remarks on the Data-Driven/Hypothesis-Driven Opposition. At the end of phase 2, entities still in the universe of hypotheses are supposed to play a causal role in the phenotype of interest. In phase 3, causal roles are 'strongly' validated. This is the 'hypothesis-driven' phase. In the philosophy of (molecular) biology, the received view states that 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. There are several rational paths that lead to the discovery of mechanisms (Bechtel and Richardson 2010; Craver and Darden 2013). Although this is not the place to review all the traditional approaches to discovering mechanisms, it should be emphasized that Weinberg and other 'traditional' molecular biologists clearly refer to these methodologies as the ones used in 'traditional' molecular biology. Phases 1 and 2 can be included in the 'mechanistic' perspective, in the sense that DDHD employs a discovery procedure that is compatible with the ones depicted by mechanistic philosophy.

Consider, for instance, the crucial distinction made by Bechtel and Richardson (2010) between localization and decomposition. These two strategies are considered as starting points in mechanistic discovery. Decomposition "assumes that one activity of a whole system is the product of a set of subordinated functions" (2010, 23), while localization tries to identify the entities that may play the subordinated functions. Clearly, there can be interplay between the two strategies. In DDHD the particular problem x and the kind of solutions implied by the background assumption $y1, y2, \ldots, yn$ lead to the assumption that the system investigated is somehow decomposable into subordinated functions. Accordingly, the system is dismantled into several

6. These genes have functions that, so far, are supposed to have nothing to do with cancer.

https://doi.org/10.1086/680332 Published online by Cambridge University Press

subcomponents *z1*, *z2*, . . . , *zn* (e.g., SNPs in a GWAS, somatic mutations in cancer genomics) that are supposed to be responsible for the subordinated functions, whatever these are. After the eliminative steps, some of the *z*'s are retained as strongly associated with the phenomenon under investigation. Most importantly, during eliminative steps, some hypotheses are also developed. If in phase 1 hypotheses take the abstract form of 'the entity *x* has a causal role in the phenotype *y*', the particular causal role is not specified at all. The strategies illustrated in section 2.2 not only eliminate spurious associations but also provide a provisional idea of the causal role.

The interplay between decomposition and localization might be conceived in parallel to some remarks made by Craver and Darden (2013) in their chapter 5. The idea is that in discovering mechanisms one looks immediately for entities or activities that might be involved in the phenomenon of interest. But the task of discovering mechanisms starts with a preliminary characterization of the phenomenon (precipitating conditions, modulating conditions, etc.). From this preliminary idea about a phenomenon, one subdivides a system into parts, identifying some as relevant. This is exactly the same with DDHD. In light of the preliminary characterization of a phenomenon *x* and the kind of solutions implied by the background assumption *y1*, *y2*, . . . , *yn* (where *x* and *y*'s 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 *z*'s are retained as being relevant parts of the mechanisms producing the phenomenon.

The gold standard to corroborate the hypothesis '*z* has the causal role *y* in the phenomenon' is a mechanistic description of how *z* is implicated in the phenomenon (for exceptions, see Boniolo 2013). In order to develop and corroborate such descriptions, molecular biology makes extensive use of experimental approaches, for example, intervening on a specific component of a system (Craver and Darden 2013, especially chap. 8). By observing the consequences of the intervention on an entity, its contribution to the whole system might be inferred.

In molecular biology, elucidating mechanisms provide both explanation and prediction. In most cases, explanation and description are equated. The idea is that "to give a description of a mechanism for a phenomenon is to explain that phenomenon, i.e. to explain how it was produced" (Machamer, Darden, and Craver 2000, 3). Moreover, knowing how a phenomenon is produced might provide clues on (1) what we should expect if we modify one of the components of the mechanisms or (2) under which conditions we should expect that the mechanisms would be in place (i.e., prediction). Most importantly, by providing descriptions of mechanisms, we have also control, that is, the ability to modify a system (Craver and Darden 2013).

Especially in contemporary molecular biology and recent biomedical attempts, the ideal of control is fundamental. This is why data-driven is complemented with hypothesis-driven: correlations and associations are not enough. Causal knowledge is the aim of phase 3.

In GWASs, the work of most screenings ends with phase 2. However, practitioners are aware 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 obtaining "functional confirmation that the variants implicated are truly causal" (McCarthy et al. 2008, 366) and reconstructing the molecular and physiological mechanisms are crucial steps. There are studies (e.g., Pomerantz et al. 2009, 2010) that select SNPs associated with a disease in many GWASs and try to fully achieve phase 3. Therefore, in GWASs practitioners try to elaborate and corroborate mechanistic descriptions of the same kind as the ones required by 'traditional' molecular biology.

Similar considerations may be drawn for cancer genomics. After discovering in phase 2 mutations or genes likely to be drivers in cancer development, practitioners elaborate mechanistic descriptions of how these entities are actually driving disease development.

Therefore, now it should be uncontroversial that DDHD is compatible with the epistemic perspective embraced by mechanistic philosophers. I am tempted to say that the discovery strategies employed in DDHD are merely a particularly interesting version of mechanistic philosophy. The interesting part is that DDHD provides a set of mechanical procedures to go through decomposition and localization. While DDHD promotes efficiency, its approach does not deviate from the general guidelines prescribed 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 illustrated by 'mechanistic philosophers', and if 'mechanistic philosophy' identifies the research strategies employed also in traditional molecular biology, then cancer genomics and GWASs (taken to exemplify the 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.

To conclude, the proposals that (1) there is not an opposition between data-driven and hypothesis-driven approaches and (2) in contemporary biology there is a hybrid of the two (DDHD) are corroborated by actual scientific practices. The hybrid is composed of the practices of generating big

data sets (by means of sequencing technologies) that are then analyzed in light of the discovery strategies typical of molecular biology.

3. The Role of Mining Studies in Contemporary Biology. The approach described in the previous section (see table 1) disentangles the scientific controversy mentioned in the introduction by showing how data-driven and hypothesis-driven approaches form a hybrid that is compatible with the received view of mechanistic philosophy. However, there are plenty of data-driven studies that cannot be reduced to DDHD. The studies I am talking about most often emerge from big consortia such as the Cancer Genome Atlas, the ENCODE project, or the 1000 Genomes Project. By joining forces, big consortia are able to generate far more data than a single scientific lab. For example, the Cancer Genome Atlas has sequenced, so far, 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 implied by its name. ENCODE has recently characterized the biochemical activities along the human genome's regions of several human cell lines (ENCODE Project Consortium 2012). Databases store these data sets, and recently computer scientists have started to look for patterns in them. It is common to find in top journals such as *Nature* or *Science* articles characterizing trends and patterns found in vast data sets. Terms such as 'comparative analyses', 'system-level characterizations', and 'emerging landscapes' have become keywords. I call these screenings "mining studies." For instance, by analyzing 3,299 tumors from 12 cancer types, Ciriello et al. (2013) discover (1) a trend that divides tumors into two classes, one characterized by somatic mutations and the other by copy-number variations, and (2) that within each major class there are specific oncogenic pathways altered. In fact, the aim of the paper is to reduce the complexity of thousands of molecular alterations discovered in thousands of tumors to a few hundred types and patterns and to categorize tumors on this basis. Some mining studies focus specif-

TABLE 1. Summary of the Hierarchy of Practices in DDHD with the Examples of GWASs and Cancer Genomics

Practice	GWASs	Cancer Genomics
Background assumptions	Disease variant hypothesis SNP hypothesis	Role of (driver) mutations in the development of cancer
Prioritization	Statistical analysis, technical derivation, replication, and comparison with ENCODE data	BMR analysis of mutations and genes
Hypothesis validation	Traditional strategies for the discovery of mechanisms	Traditional strategies for the discovery of mechanisms

https://doi.org/10.1086/680332 Published online by Cambridge University Press

ically on copy-number variations (Li et al. 2012; Kim et al. 2013; Zack et al. 2013) or on somatic mutations (Kandoth et al. 2013), while others focus on the analysis of trends in the functional annotations of the human genome (ENCODE Project Consortium 2012). Despite the increasing number of these studies, their purpose is not clear.

3.1. The Structure of Mining Studies. The structure of these studies is straightforward. In mining studies, computer scientists look for associations of metadata. Metadata are labels 'attached' to a particular bit of data (or to a data set), and they are used to describe the bits of data themselves. In other words, the metadata define what the data are about. The idea of mining studies is that researchers look for robust regularities in metadata associations. For example, Kim et al. (2013) mine the Cancer Genome Atlas database, looking for a particular genomic structural rearrangement called copy-number variation. However, they do not look for regularities of this type of genomic rearrangement with respect to their position along the human genome, but rather with respect to another metadatum, that is, tumor type (defined by tissue of origin). Then, for each set of genomic rearrangements in each tumor type, Kim and colleagues look for the genes located within the region amplified or deleted. By doing this, it is possible also to capture certain regularities and to identify, for each tumor type, the biological processes that one might expect to find disrupted. In other words, it is possible to formulate 'generalizations' such as 'in lung cancer, copy-number variations deregulate the pathways *x*, *y*, and *z*'.

Therefore, the structure of mining studies is simple:

1. Scientists look for associations between different metadata labels in order to uncover macroregularities.
2. Macroregularities are, strictly speaking, predictions in the sense that they provide an expectation of what is likely to be observed in similar contexts.

3.2. What Mining Studies Are Not

3.2.1. Mining Studies Are Not Driven by Eliminative Induction, and Background Assumptions Are Weaker than in DDHD. Mining studies are regarded as instances of data-driven research, but they share few features with DDHD. What DDHD and mining studies have in common is, first, the use of background assumptions. However, in mining studies background assumptions play a substantially weaker role than in DDHD. Background assumptions of mining studies include

1. the theoretical basis of the computational tools used to identify associations;

https://doi.org/10.1086/680332 Published online by Cambridge University Press

2. the fact that pattern discovery is metadata-laden,[7] that is, it is possible to find associations only within the categories of a preexisting taxonomy system *x* (e.g., gene ontology).

Background assumptions in mining studies provide weaker guidelines than DDHD. There is not a clear scientific problem aside from 'which are the interesting regularities in this data set?' Moreover, the tentative answer to this broad problem is not sufficiently specific, as it is merely the idea that the regularities to be discovered depend strictly on a sort of power set of all possible ways of associating the metadata labels. It is fair to say that this power set represents a sort of initial universe of hypotheses for mining studies. However, the universe of hypotheses in mining studies is narrowed differently than in DDHD. In DDHD, adding additional assumptions narrows the initial universe. In mining studies, eliminative principles are weaker and more vague. They are based on whether a computer scientist considers a pattern sufficiently robust to be defined as regularity. Apart from these, there are no other background assumptions constraining the discovery of patterns.

3.2.2. Mining Studies Are Not Explanatory. As has been suggested, DDHD aims to provide acceptable hypotheses. The way hypotheses are taken to be acceptable is by discovering, through experimental manipulations, that the entities hypothesized to play a causal role in a phenomenon are actually embedded in a mechanism. The gold standard is a description of a mechanism where one or more of the entities that survived eliminative induction play a key role.

In mining studies, the ideal of discovering mechanisms plays no role. What mining studies provide are predictions or generalization. While it is true that providing a mechanistic explanation enables the formulation of predictions, the reverse is clearly false. The fact that an SNP has 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, merely finding the correlation between an SNP and diabetes does not provide mechanistic descriptions that can explain the association. Douglas (2009) argues that the relation between explanation and prediction is a functional one. Actually, "explanations provide the cognitive path to predictions, which then serve to test and refine the explanation"

7. To better understand what 'metadata-laden' means, imagine that the taxonomical system used by Ciriello et al. to embed their research classifies 'DNA mismatch repair' and 'p-53 mediated apoptosis' under the same label (see fig. 1). 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 (see fig. 1) because we would classify 'DNA mismatch repair' and 'p-53 mediated apoptosis' as the same phenomenon.

(Douglas 2009, 454). In a nutshell, predictions are valuable because they force us to test our explanations.

In the case of DDHD, the reverse is true: predictions (the prioritized hypotheses) provide the cognitive path to mechanistic explanation. This can be true also for mining studies. The association of ovarian cancer with alteration of p-53 mediated apoptosis is a prediction suggesting an experimental path to elaborate mechanistic descriptions. However, what I argue in the next section is that the predictions established by mining studies do not suggest directly a path to uncover mechanisms, but rather something subtler.

3.3. Predictions Established by Mining Studies Are Generalizations Forming Additional Background Assumptions for DDHD. In section 3.2 I have shown that the role of mining studies in contemporary biological research is not entirely clear. What is the function of their predictions in contemporary biological research?

Predictions may also be seen 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 1997, S477). For instance, it is possible to generalize the association between two transcription factors (x and y) through the ENCODE data so that whenever one finds x, y is likely to be found. Another example is figure 1 (taken from Ciriello et al. 2013), where one may observe a strong association between the gene *PIK3CA* and breast cancer. This means that there is a high probability of finding *PIK3CA* mutated in breast cancer.

However, generalizations uncovered by mining studies play, in my opinion, a subtler role than traditional predictions or expectations. My claim is that generalizations inferred from mining studies might provide some of the eliminative principles used to narrow the universe of hypotheses elaborated in phases 1 and 2 of DDHD. Let us see how.

Consider GWASs. Above, I have said that one late eliminative step is to elaborate a preliminary functional characterization of SNPs by looking at ENCODE data (Germain et al. 2014). If an SNP either does not overlap a functional region or overlaps with a region whose function is not related to the phenotype of interest, then the SNP is eliminated from the universe of hypotheses. ENCODE and other big projects' annotations are clustered with other functional annotations from other big projects in a database called RegulomeDB (Boyle et al. 2012; Schaub et al. 2012).[8] The aim of this database is to provide comprehensive generalizations of the biochemical activities on the human genome, as well as a quantification of the confidence that a particular genomic region is likely to engage in a particular activity. This

8. See http://www.regulomedb.org/.

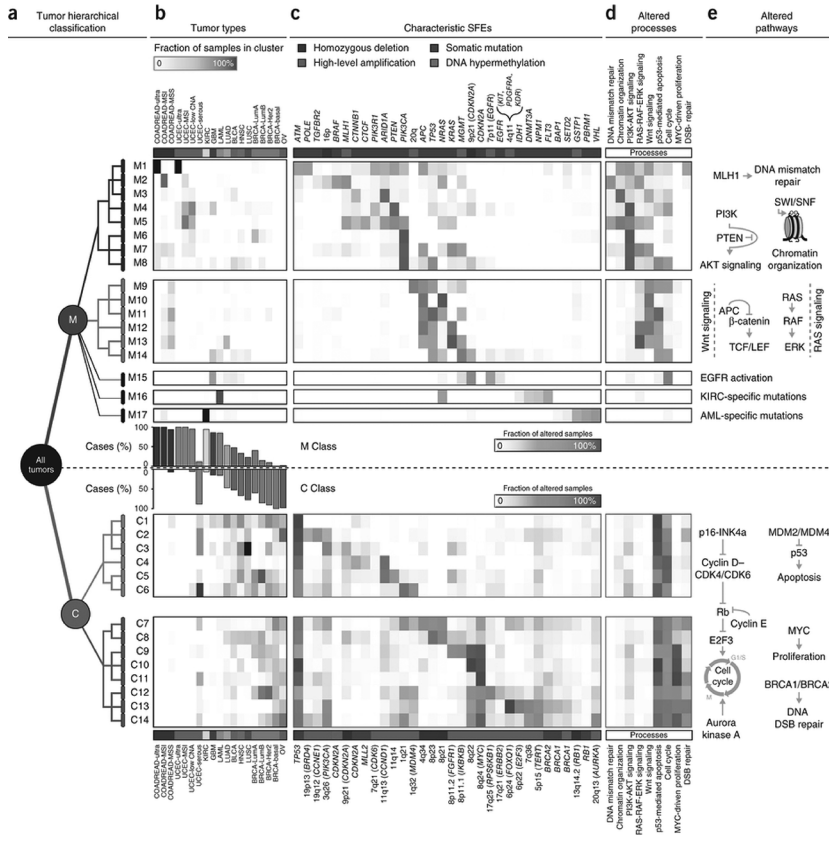

Figure 1. Visual summary of the patterns identified in the mining study (Ciriello et al. 2013). *a*, Tumors are divided into two main classes: mutational tumors (M) and copy-number tumors (C). For M, Ciriello and colleagues have identified 17 subclasses, while within C they have found 14 distinct 'oncogenic signatures'. *b*, Each tumor type (defined by tissue of origin) falls, mostly, in one of the subclasses. *c*, Each tumor type has specific functional alterations related to specific genes. *d,e*, Each tumor type has specific cellular processes and pathways disrupted. Shades of gray represent the fraction of samples in the cluster. The figure is taken from Ciriello et al. (2013) and has been modified. In particular, it has been transformed into black and white. Enlarged figure available as an online enhancement.

type of classification can be seen as a generalization of the biochemical activity happening in each region of the human genome in the form of 'the genomic region *x* is biochemically active in such and such a way'. These generalizations suggest useful principles to be used in order to interpret SNPs and hence to narrow the universe of hypotheses of GWASs. For instance, imagine that

https://doi.org/10.1086/680332 Published online by Cambridge University Press

the SNP *x* falls in a region *y* that is associated with a transcription factor binding site influencing a particular gene *z*. One might speculate that *x* has a small effect on the phenotype of interest because it lies in *y*, which in turn regulates the expression of *z*. By means of generalizations provided by databases like RegulomeDB about *y*, biologists can either eliminate SNPs or develop a more precise idea of the causal role that an SNP might have.

Similar considerations may be drawn for cancer genomics. As Raphael et al. (2014) emphasize, a challenge in cancer genomics is to identify driver mutations and to understand their effects on pathways and cellular processes. The idea is that certain genes (in light of functions and the pathways genes participate in) might be good proxies for driver mutations. As emphasized in the literature (Vandin et al. 2012; Raphael et al. 2014), 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? There is a sort of 'store' of relevant pathways in cancer genomics. This is represented in a propositional form by famous reviews (Hanahan and Weinberg 2011; Garraway and Lander 2013; Vogelstein et al. 2013) or by textbooks. However, mining studies have started to provide such a 'store' in a more comprehensive way. Consider the mining study of Ciriello et al. (2013). By means of their generalizations, in DDHD it is possible to compile a list of cancer-specific pathways to check during the narrowing of the universe of hypotheses. Let us see this through a specific example. Consider figure 1.

This figure represents a visual summary of the patterns identified in Ciriello et al. (2013). Actually, this figure is a representation of the mining study. Imagine that a physician has a patient affected by colon cancer (ultramutator variety). The physician then decides to genotype the tumor of the patient in order to better understand the genomic features of the tumor. A guide on what to look for in the genome is provided by figure 1. First, if the tumor is colon cancer ultramutator (column *b*, COADREAD-ultra) variety, then it is an M1 tumor (column *a*). Second, somatic driver mutations should be located in selected genes such as *ATM*, *APC*, or *PTEN* and other genes that are located in the same pathways (column *c*). Therefore, the physician will observe only certain genes and not others (column *c*). Moreover, in order to identify disrupted pathways and altered cellular processes, the study of Ciriello et al. is a source of criteria for prioritization. In the case of colon cancer, disrupted processes and pathways include chromatin organization, PI3K-AKT signaling, and so on, as shown in columns *d* and *e*.

To sum up, the associations found by Ciriello et al. might be considered as a set of predetermined genes and pathways that any researcher should compare with her list of mutations, genes, and copy-number variations. If certain genes do not overlap with this set, 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.

3.4. Mining Studies as Instances of Exploratory Experimentations. In the previous section I argued that mining studies elaborate generalizations aimed at creating or complementing eliminative principles for DDHD. 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. Interestingly, mining studies might be considered as a peculiar case of exploratory experiments. Actually, mining studies meet many of the features of exploratory experiments ascribed by Steinle (1997). 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, so-called big data biology is still in its infancy, and only recently have scientists started to uncover preliminary generalizations. Steinle also adds that exploratory experiments are not theory-free but rather 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 for mining studies and explorations of genomes. As is now widely shared consensus, 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 for mining studies.

4. Conclusion. In this paper I have tried to make sense of the recent debate on the nature of big data studies in molecular biology.

First, I considered a recent proposal claiming that data-driven studies in biology are hybridized with traditional methodologies of mechanistic discovery. I provided a framework composed of three stages where data-driven methodologies are actually included in the broader mechanistic perspective. The aim of the first part of this article was to show that the controversy between scientists supporting 'traditional' molecular biology ('hypothesis-driven') and scientists supporting biology 'post–Human Genome Project' ('data-driven') is ill posed. The only difference between DDHD and tradi-

https://doi.org/10.1086/680332 Published online by Cambridge University Press

tional methodologies of mechanistic discovery is that the former provide quasi-algorithmic procedures to follow the strategies of decomposition and localization (Bechtel and Richardson 2010), that is, to identify components that are supposed to be involved in the phenomenon of interest.

The second aim of this work was to find a place in contemporary biology for what I call 'mining studies'. These studies are 'data-driven', but they seem not to be instances of DDHD in any straightforward way. However, mining studies aim to elaborate generalization that can be used to provide guidelines for the formulation of eliminative principles to be used in DDHD research. In particular, mining studies play the role of exploratory experiments in navigating the immense sea of data generated by contemporary sequencing projects. It is not unfair to say that mining studies represent a sort of 'store' for discovery à la Darden (Craver and Darden 2013). These stores not only involve entities and activities but also provide important hints about entities' involvement in various phenomena.

Therefore, it seems that the discovery strategies of DDHD, although they do not deviate from the ones described by mechanistic philosophy, crucially make use of exploratory experimentations. Mining studies can be seen as one type of heuristic strategy fitting the general mechanistic perspective.

Molecular biology is undergoing a deep change, especially after the Human Genome Project. Examples include the combination of multiple types of expertise, the increasing importance of computer scientists, and the extensive use of biological databases even for elementary purposes. However, from my analysis it seems that the way biological systems are analyzed has not changed as dramatically as other aspects related to biological research. Vast data sets are useful mainly because, when analyzed by means of powerful computational resources, they are taken as more comprehensive starting points for typical decomposition and localization strategies. The power of 'data-driven' studies (including many 'system' biology' studies; see Gross 2013) lies exactly in the amplitude of data provided. As Dulbecco anticipated many years ago (1986), having a higher system-level view of all the biological entities (genes, entities, etc.) that in principle might be involved in the development of a phenotype can be a more effective starting point to discovery than a piecemeal, tried-and-tested approach, where one picks up entities to analyze randomly. The same applies for mining studies. Within the traditional discovery strategies of molecular biology, the power of mining studies lies in the hints they provide for the prioritization of certain entities.

REFERENCES

Alberts, Bruce. 2012. "The End of 'Small Science'?" *Science* 337 (6102): 1583.

Bechtel, William, and Robert C. Richardson. 2010. *Discovering Complexity—Decomposition and Localization as Strategies in Scientific Research*. Cambridge, MA: MIT Press.

Boniolo, Giovanni. 2013. "On Molecular Mechanisms and Contexts of Physical Explanation." *Biological Theory* 7 (3): 256–65.
Boyle, Alan P., et al. 2012. "Annotation of Functional Variation in Personal Genomes Using RegulomeDB." *Genome Research* 22 (9): 1790–97.
Brenner, Sydney. 1999. "Syllicon Valley Fever." *Current Biology* 9 (18): R671.
Brookfield, John F. Y. 2010. "Q&A: Promise and Pitfalls of Genome-Wide Association Studies." *BMC Biology* 8:41.
Ciriello, Giovanni, Martin L. Miller, Bulent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. 2013. "Emerging Landscape of Oncogenic Signatures across Human Cancers." *Nature Genetics* 45 (10): 1127–33.
Craver, Carl F., and Lindley Darden. 2013. *In Search of Mechanisms*. Chicago: University of Chicago Press.
Douglas, Heather E. 2009. "Reintroducing Prediction to Explanation." *Philosophy of Science Science* 76 (4): 444–63.
Dulbecco, Renato. 1986. "A Turning Point in Cancer Research: Sequencing the Human Genome." *Science* 231 (4742): 1055–56.
Earman, John. 1992. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.
The ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414): 57–74.
Forber, Patrick. 2011. "Reconceiving Eliminative Inference." *Philosophy of Science* 78 (2): 185–208.
Garraway, Levi, and Eric Lander. 2013. "Lessons from the Cancer Genome." *Cell* 153 (1): 17–37.
Germain, Pierre Luc, Emanuele Ratti, and Federico Boem. 2014. "Junk or Functional DNA? ENCODE and the Function Controversy." *Biology and Philosophy* 29:807–31.
Golub, T. 2010. "Counterpoint: Data First." *Nature* 464 (7289): 679.
Gross, Fridolin. 2013. "The Sum of the Parts: Heuristic Strategies in Systems Biology." PhD diss., University of Milan.
Guessous, Idris, Marta Gwinn, and Muin J. Khoury. 2009. "Genome-Wide Association Studies in Pharmacogenomics: Untapped Potential for Translation." *Genome Medicine* 1 (4): 46.
Hanahan, Douglas, and Robert Weinberg. 2011. "Hallmarks of Cancer: The Next Generation." *Cell* 144 (5): 646–74.
Hawthorne, James. 1993. "Bayesian Induction Is Eliminative Induction." *Philosophical Topics* 21 (1): 99–138.
Hunter, David J., David Altshuler, and Daniel Rader. 2008. "From Darwin's Finches to Canaries in the Coal Mine—Mining the Genome for the New Biology." *New England Journal of Medicine* 358:26.
Kandoth, Cyriac, et al. 2013. "Mutational Landscape and Significance across 12 Major Cancer Types." *Nature* 502 (7471): 333–39.
Keating, Peter, and Alberto Cambrosio. 2012. "Too Many Numbers: Microarrays in Clinical Cancer Research." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 37–51.
Kell, D. B., and S. G. Oliver. 2003. "Here Is the Evidence, Now What Is the Hypothesis? The Complementary Roles of Inductive and Hypothesis-Driven Science in the Post-genomic Era." BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology 26 (1): 99–105.
Kim, Tae-Min, Ruibin Xi, Lovelace Luquette, Richard Park, Mark Johnson, and Peter J. Park. 2013. "Functional Genomic Analysis of Chromosomal Aberrations in a Compendium of 8000 Cancer Genomes." *Genome Research* 23 (2): 217–27.
Kitcher, Philip S. 1993. *The Advancement of Science*. New York: Oxford University Press.
Kitsios, Georgios, and Elias Zintzaras. 2009. "Genome-Wide Association Studies: Hypothesis-Free or 'Engaged'?" *Translational Research* 154 (4): 161–64.
Lawrence, Michael, et al. 2013. "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes." *Nature* 499 (7457): 214–18.
Leonelli, Sabina. 2012a. "Classificatory Theories in Data-Intensive Science." *International Studies in the Philosophy of Science* 26 (1): 47–65.

———. 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.

Li, Yudong, Li Zhang, Robyn Ball, Xinle Liang, Jianrong Li, Zhenguo Lin, and Han Liang. 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.

Machamer, Peter, Lindey Darden, and Carl Craver. 2000. "Thinking about Mechanisms." *Philosophy of Science* 67:1–25.

McCarthy, Mark I., Goncalo R. Abecasis, Lon R. Cardon, David Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. 2008. "Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges." *Nature Reviews Genetics* 9 (5): 356–69.

Mitchell, Sandra D. 1997. "Pragmatic Laws." *Philosophy of Science* 64 (Proceedings): S468–S479.

Norton, John. 1995. "Eliminative Induction as a Method of Discovery: How Einstein Discovered General Relativity." *The Creation of Ideas in Physics*, ed. Jarrett Leplin, 29–69. Dordrecht: Kluwer.

O'Malley, Maureen. 2007. "Exploratory Experimentation and Scientific Practice: Metagenomics and the Proteorhodopsin Case." *History and Philosophy of the Life Sciences* 29 (3): 335–58.

O'Malley, Maureen, and Orkun S. Soyer. 2012. "The Roles of Integration in Molecular Systems Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 58–68.

Peschard, Isabelle, and Bas van Fraassen. 2014. "Making the Abstract Complete: The Role of Norms and Values in Experimental Modelling." *Studies in History and Philosophy of Science* A 46:3–10.

Platt, John R. 1964. "Strong Inference." *Science* 146 (3642): 347–53.

Pomerantz, M. M., et al. 2009. "The 8q24 Cancer Risk Variant rs6983267 Shows Long-Range Interaction with MYC in Colorectal Cancer." *Nature Genetics* 41 (8): 882–84.

———. 2010. "Analysis of the 10q11 Cancer Risk Locus Implicates MSMB and NCOA4 in Human Prostate Tumorigenesis." *PLoS Genetics* 6 (11): e1001204.

Raphael, Benjamin J., Jason R. Dobson, Layla Oesper, and Fabio Vandin. 2014. "Identifying Driver Mutations in Sequenced Cancer Genomes: Computational Approaches to Enable Precision Medicine." *Genome Medicine* 6 (1): 5.

Rheinberger, H.-J. 2011. "Infra-experimentality: From Traces to Data, from Data to Patterning Facts." *History of Science* 49 (337).

Schaub, Marc, Alan P. Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. 2012. "Linking Disease Associations with Regulatory Information in the Human Genome." *Genome Research* 22 (9): 1748–59.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310.

Smalheiser, Neil R. 2002. "Informatics and Hypothesis-Driven Research." *EMBO Reports* 3 (8): 702.

Steinle, Friedrich. 1997. "Entering New Fields: Exploratory Uses of Experimentation." *Philosophy of Science* 64 (Proceedings): S64–S74.

Strasser, Bruno. 2011. "The Experimenter's Museum—GenBank, Natural History, and the Moral Economies of Biomedicine." *Isis* 102 (1): 60–96.

Vandin, Fabio, Eli Upfal, and Benjamin J. Raphael. 2012. "Finding Driver Pathways in Cancer: Models and Algorithms." *Algorithms for Molecular Biology: AMB* 7 (1): 23.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz Jr., and Kenneth W. Kinzler. 2013. "Cancer Genome Landscapes." *Science* 339 (6127): 1546–58.

Waters, C. Kenneth. 2007. "The Nature and Context of Exploratory Experimentation." *History and Philosophy of the Life Sciences* 29:1–9.

Weinberg, Robert. 2010. "Point: Hypotheses First." *Nature* 464 (7289): 678.

Zack, Travis I., et al. 2013. "Pan-cancer Patterns of Somatic Copy Number Alteration." *Nature Genetics* 45 (10): 1134–40.