Penultimate draft. Accepted for publication in *Perspectives on Science*. Please cite published version

BEYOND 'TRAPPED PETS' AND 'RED BUTTONS': BIOINFORMATICS AS AN EXPERIMENTAL DISCIPLINE

Emanuele Ratti[1], Department of Philosophy, University of Bristol

Giuseppe D'Agostino, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

Abstract. The past few years have witnessed a growth in the interest on the historical and philosophical dimensions of bioinformatics as a discipline. Despite the importance of bioinformatics in addressing the issues raised by the growing amount of biological data, data management is often seen as all it has to offer to biology. However, the emphasis on data management may come at the expense of understanding how bioinformatics generates genuine biological knowledge beyond its instrumental value for bench biologists. Some authors have moved the first steps beyond data management, and towards the characterization of bioinformatics as a unique epistemic endeavor, by stressing how its experimental practices can be conducive to biological knowledge. In this article, we build upon these attempts, and by using a detailed case study from the field of single cell transcriptomics (i.e., RNA velocity), we provide a fully-fledged characterization of bioinformatics as an experimental discipline.

Keywords: bioinformatics; big data biology; experiments; molecular biology

1. INTRODUCTION

Computers have been used since the 1950s in various biological contexts, including molecular biology. Key pioneering figures include people such as Robert Ledley, Joshua Lederberg, and Walter Goad, who attempted to develop simulations and data modeling techniques to answer biological questions (November 2012; Stevens 2013; Strasser 2017). However, as documented by Stevens (2013) and Strasser (2017), many molecular biologists were initially reluctant to use computers, deeming them pointless and thus hindering early attempts at introducing computational projects in biology. But since the *data deluge* started in the early 1980s, computational assets such as databases started to attract the interest of molecular biologists, to the extent that a new discipline called 'bioinformatics'[2] has slowly emerged. While the data

[1] Corresponding author, mnl.ratti@gmail.com

[2] The term "bioinformatics" is used here as a stand-in for all the other definitions used in the community: computational biology, systems biology, etc. We are aware that the community does not always view these terms interchangeably, with some considering "bioinformatics" as the aspects merely related to software engineering (thus pertaining more to computer scientists), and "computational biology" as a more rounded way of studying biology using computational tools and methods (see https://www.kennedykrieger.org/sites/default/files/library/documents/research/center-labs-

management dimension has played a central role, the development of early computational projects based on software and tools to model biological data has nonetheless continued in parallel. However, historical investigations have focused especially on the data management and sequence analysis aspects of bioinformatics, and to our knowledge, a detailed history of other dimensions of bioinformatics has yet to be written.

The situation is slightly better in philosophy of science. Philosophical works engaging with the epistemology of computationally-driven disciplines such as genomics have only very recently started to discuss bioinformatics as a discipline *per se*, without implicitly assuming that it is just a set of tools and solutions to store and sometimes analyze data. In particular, recent works by Leonelli (2016; 2019), Strasser (2017), and Stevens (2013), while describing the epistemic ramifications of data management practices in the biological context in great detail, have also attempted to go a step beyond the view of bioinformatics as data management and automated analysis, by emphasizing the experimental dimension[3] and the proper goals of this discipline (see Appendix 1 for an overview of these positions). By building on these attempts, the goal of this article is to develop a full-blown account of bioinformatics as an endeavor with its own epistemic goals. In our work, bioinformatics is understood as a discipline that, by engaging in experimental activities with virtual experimental systems (whose origin is nonetheless material) through the development of new computational[4] tools, generates new kinds of data that wet-lab biologists cannot create. Our account coherently integrates the aspects described by Stevens, Leonelli, and Strasser, with novel facets of bioinformatics practices in the molecular biological context. By doing so, we shed new light on bioinformatics as a unique epistemic and experimental culture.

1.1 Motivations

The motivations for the present work are two.

One motivation is to fill a gap in the philosophical, historical and sociological literature on the nature of bioinformatics. This has been often seen in relation to other biological disciplines, with only recent works trying to characterize bioinformatics in its own right (as

cores/bioinformatics/bioinformatics-def.pdf). They are, however, often used interchangeably especially in the context of multidisciplinary laboratories and collaborations. We acknowledge that the differences underlying these terms can be relevant, but for the sake of simplicity and in keeping with the dynamics in multidisciplinary environments we will use one umbrella term.

[3] An isolated case in which the experimental dimension of bioinformatics is (Boem and Ratti 2016)

[4] 'Computational' here is synonymous with 'data-intensive', 'Big Data', or 'AI': they are different words to refer to the same class of tools

discussed in Appendix 1). By building on these attempts, this paper should be seen as contributing to the philosophical, historical, and sociological understanding of a discipline that has been often considered *ancilla biologiae*, rather than a proper subfield of biology.

The second motivation comes from an unfavorable situation in which many bioinformaticians work, which, to be improved, requires a richer epistemic account of bioinformatics practices. In particular, we refer to what has been called 'the trapped bioinformatician[5]' or 'the pet bioinformatician[6]' syndrome. There is a tendency in laboratory-based groups to consider bioinformaticians as valuable resources to manage, curate, and analyze the data that "wet-lab" biologists generate, but not so much as pursuers of genuine biological questions themselves. This implies that bioinformaticians will have difficulties in completing their own projects, which often require bench biologists to reciprocate the time that bioinformatics practitioners have spent in analyzing data belonging to wet-lab projects. This culture reflects a view of bioinformatics practice as mere 'red button-pushers' that initiate automated analysis procedures[7]. This situation, we claim, generates a divide. On the one hand, wet-lab biologists think about bioinformatics mostly in instrumental terms as data management and analysis, while bioinformaticians feel that they generate genuine biological knowledge themselves. On the other hand, wet-lab biologists who generate large amounts of data in high throughput experiments often lack the expertise to analyze such datasets, thus having to rely on bioinformaticians for important steps, decisions and biological interpretation of these experiments, but the nature of such decisions and its impact is underestimated. We call this divide *epistemic alienation*: bioinformaticians generate genuine biological knowledge, but they are excluded from the intellectual category of 'knowledge makers'; at the same time, wet-lab biologists cannot make sense of important parts of their own work as they need bioinformaticians' inputs to interpret their own experiments. The divide, which implies a subordination of bioinformaticians to wet-lab biologists, has been well documented by the massive, decade-long STS study of bioinformatics culture by Andrew Bartlett, Bart Penders, Jamie Lewis, and others (Lewis and Bartlett 2013; Bartlett et al 2016; Lewis et al 2016; Bartlett et al 2017). One highlight of their studies is that "many view bioinformatics as a 'service', rather than a scientific field in its own right (...) [this] renders the intellectual contribution of bioinformaticians invisible, hidden in the 'black-box'" (Bartlett et al 2017, p 2). One

[5] http://davidsdatablog.blogspot.com/2018/12/trapping-pet-bioinformatician-for-lab.html
[6] http://www.opiniomics.org/a-guide-for-the-lonely-bioinformatician/
[7] See for instance this ironic post by Torsten Seeman
https://x.com/torstenseemann/status/433448248921956352?prefetchTimestamp=1732106085400

consequence of this view is a shady distribution of credits among wet-lab scientists and bioinformaticians. The results of this study are supported by an impressive variety of empirical evidence[8], corroborated by more insights (Markowetz 2017; Grabowski and Rappsilber 2019; Way et al 2021), though with recent slight improvements (Calder et al 2021).

Epistemic alienation has a strong sociological and political component. As Bartlett et al (2017) say, science is necessarily tied to "institutional and organizational arrangements" (p 2) which shape power dynamics. From this point of view, there is not much that we can do with the present article. However, what we can do is to dismantle philosophical prejudices lying at the roots of epistemic alienation. Therefore, we argue against the prejudices that bioinformaticians do mostly data management and that their work can be increasingly automated, and most importantly that they cannot produce novel biological knowledge by working on purely computational projects. An epistemic account of bioinformatics practice can show that there is more to this discipline than just data management and automated data analysis, and that bioinformatics is indeed an experimental science, as much as molecular biology is. The emphasis on 'experimental' is essential, given the old illustrious theme in molecular biology (Strasser 2017) that those who generate genuine biological knowledge are the ones doing the experimental work.

1.2 The structure of the article

The structure of the article is as follows. In Section 2 we identify the philosophical assumptions behind epistemic alienation and the idea that bioinformatics should be subordinated to wet-lab biologists. We introduce the concept of 'epistemic driver', which designates those scientific actors leading a research project and co-opting other people's labor to achieve their own epistemic goals. We explain how in biology being an epistemic driver is strictly connected to the role of experimenting, understood as a particular kind of material intervention aimed at creating new data types or new data that are indications of biological phenomena. But, a popular view claims, bioinformaticians do not do that: they only manage data and initiate automated procedures. What impedes bioinformaticians to be epistemic drivers is then a twofold problem: they do not do experiments, and they do not have material access to phenomena. In Section 3, we delineate in detail the case of RNA velocity as a paradigmatic example of bioinformatics experimentation, by showing how the model and data type of RNA

[8] This includes including content-analysis of bioinformatics articles, ethnographic fieldwork, interviews of almost 100 bioinformatics, and a survey of 300 bioinformaticians

velocity is actually discovered by intervening in various formal ways on data that have been 'converted' from 'real-world data'. In Section 4 we develop our own account of bioinformatics as an experimental practice by showing in which sense cases like RNA velocity are instances of biological experimentation, and by distinguishing two types of 'experiments' in bioinformatics that we call 'strong' and 'hard'. Finally, in Section 5, we address the concern about the missing materiality of bioinformatics experimentation. All in all, this will show that bioinformaticians can be epistemic drivers.

2. PHILOSOPHICAL ASSUMPTIONS BEHIND EPISTEMIC ALIENATION

Our starting point is the notion of 'epistemic driver'. We define an epistemic driver in a scientific group as an individual who, in leading a research project, produces scientific knowledge and co-opts other individuals' expertise to achieve his/her own epistemic goals. An epistemic driver controls the unfolding of a scientific project. This is akin to making 'path-dependent' decisions that ended up framing the general discovery strategy of a scientific project. Concretely, this means deciding the experiments to perform, how results should be interpreted, and how the efforts of other individuals should be allocated to achieve an epistemic goal that he/she chooses. Furthermore, this is also going to influence the 'story' or the 'narrative' that will be written in scientific articles[9]. When we argue that epistemic drivers are 'co-opting' other people's work, we are not saying that they force other individuals. As we will see, in biological labs there are different projects, and hence different epistemic drivers, and by offering one's own services for another project, reciprocity is expected (Knorr-Cetina 1999). A realistic picture is that, within each laboratory, there is an intricate network of projects and hence of epistemic drivers. It is possible to 'zoom-out' and identify groups of individuals that can be in principle epistemic drivers, and groups of individuals that cannot. For example, in a biological laboratory PhD students and postdocs usually lead their own projects, and hence have their own epistemic goals, while technicians do not. This means that PhD students or postdocs can become, at least in principle, epistemic drivers, while technicians cannot because they only provide a service to epistemic drivers. In this context, bioinformaticians have struggled to be recognised as 'epistemic drivers'.

The concept of "epistemic driver" is useful to describe general situations in experimental research, regardless of the field of study. Highly collaborative research groups -

[9] This is especially true for postdocs, and to a lesser extent for PhD students. But in all these cases, the PI has also a significant role in directing the discovery strategy, as well as deciding on the final 'narrative' (indeed, there are laboratories where PIs write all the articles),

including consortia - will have different projects where an individual (or, more rarely, a group of individuals) has a specific research question, studies the existing literature to identify relevant gaps, designs and executes experiments, interprets the data and compiles these interpretations in a communicable form such as visualizations or reports.

2.1 Epistemic Drivers in Macromolecular Biology

In order to understand how bioinformaticians may be denied the role of epistemic drivers, it is important to describe exactly in which sense traditional bench biologists can be defined as such. Three aspects must be emphasized at the onset.

First, the figure of the epistemic driver can be investigated from two perspectives. On the one hand, there is a socio-cultural point of view, emphasizing the power dynamics leading some specific professional, academic, and scientific figures to become 'driver' rather than others. A second angle concerns the 'characteristics' of epistemic drivers as such, in particular in the context of molecular biology or, to use Morange's expression (2008), *macromolecular biology*, which includes disciplines developed from the molecular vision, such as systems biology, the various -omics, etc. We are interested in the latter angle, even though there might be much to say about the former.

Second, to understand the epistemic reasons for being epistemic drivers in macromolecular biology, we have also to consider (a) the environment in which biologists work, and (b) the conditions of possibility for discovering how biological phenomena are constituted.

Let us start with (a), namely the laboratory. An important ethnographic study investigating the epistemic dimension of macromolecular biological labs is Knorr-Cetina's classic *Epistemic Cultures* (1999), which we use as a starting point. According to Knorr-Cetina, macromolecular biology is a discipline characterized by 'object-oriented processing', which is the continuous manipulation and production of material objects, such as plasmids, cell lines, etc., that are generated and used following protocols. The laboratory has a two-tier structure that is characterized by material objects. The first provides and maintains the materials necessary in a laboratory, while members of the second use the working material for experimental work in ways that are dictated by their epistemic goals.

We need to 'zoom-in' to the second layer in order to grasp (b), namely the conditions of possibility for discovering how biological phenomena are constituted, to which only certain practitioners (i.e. the epistemic drivers) have access. In this layer, there are "massive transformations [brought] to bear on objects" (p 85). In her rich description of the nature of

object-oriented processing, Knorr-Cetina emphasizes the importance of the experiences, the body, and the senses of biologists. In order to be a good biologist, one has to develop sophisticated experimental skills, which means being able to tinker with experimental systems in efficient ways. The lives of biologists are characterized by "daily interactions with material things, (…) the need to establish close relationships with the materials" (1999, p 86). Good biologists have to develop a deep personal knowledge of their own experimental systems (Rheinberger 1997). This is because protocols have to be adapted to the specificities of the materials biologists are working with, requiring an ability to 'feel' the experimental system (Keller 1983) in a way that protocols have to "be negotiated in practice with obdurate materials and living things" (Knorr Cetina 1999, p 88). In this context, a *necessary* condition for being an epistemic driver is having access to experimental systems and being able to manipulate them. In other words, the concreteness and materiality of experiments seem to play a central role.

2.1.1 Experimental activities, materiality, and epistemic drivers

Let us start with 'experiment' and 'experimental'. It is beyond the scope of this article to provide a precise account of 'experiments' – the topic and the literature would require a separate book-length treatise. What we do here is to highlight a few aspects associated with experiments that are important in this context.

There is a general way of understanding 'experimental' (Strasser 2017), which designates a "broad range of research practices, including both experimentation intended to control and experimentation intended to analyze" (p 14). Here we especially emphasize the 'intervention/manipulation' aspect of experimentation by focusing on aspects of laboratory science that "interfere with the course of that aspect of nature that is under study" (Hacking 1992, p 33). Parker (2009) characterizes experiments as investigative activities involving intervention "on a system in order to see how properties of interest of the system change", where an intervention is "an action intended to put a system into a particular state" (p 487). Knorr-Cetina captures this specificity in biology, noticing that many experiments subject "specimens to procedural manipulations (...) experiments deploy and implement a technology of intervention" (1999, pp 36-37). Rheinberger (1997) also emphasizes the intervention/manipulation dimension of experimentalists, by stressing that experimentalists (in his case, molecular biologists) are 'tinkerers' rather than engineers (1997, p 32). But just tinkering is not enough. Tinkering with biological systems is a necessary aspect of experimental activities, but one can tinker in non-experimental ways. Experimenting is (a) tinkering with a system's parts in a controlled setting, (b) recording unforeseen consequences

in order to understand parts' behavior, (c) with some biological questions in mind. Tinkering without a question or interest in mind is not experimenting, as it is not experimenting just tinkering with known consequences (like repairing a system we understand perfectly) or without controls. The mention of controls is particularly salient, because experimenting is not mindless tinkering; it needs 'confidence-building' strategies (Franklin 1986; Parker 2008), namely ways of checking whether the experimental activity is at least internally valid – this requires practices of calibration, consistency of results with known intervention or even with theory, robustness, etc.

Let us now turn to the 'concreteness' or 'material' component. Experiments defined in this way are central because the way biological phenomena are produced and/or maintained cannot be directly observed, and biologists have to find 'creative' (though reliable!) ways to 'force' experimental systems to 'reveal' something about biological phenomena. One central way in which this is done is by manipulating experimental systems in order to generate either *novel data types* or simply *new data* that can constitute evidence for phenomena (Leonelli 2016; Bogen and Woodward 1988). On this account, data are "the marks that some section of the world [i.e. in this case, specific biological phenomena] makes when it moves through some recording field" (Lowrie 2017, p 9). Biologists manipulate experimental systems in ways that will 'force' the phenomenon to leave new types of traces (especially if they want to discover something new) or just specific traces that they know are indicative of a specific phenomenon[10]. In the tradition of macromolecular biology as depicted by Knorr Cetina, these experimental activities have an important material dimension: new data types or specific traces are created by *materially* manipulating experimental systems. To take a common example, in order to study the biological phenomenon we call 'genome', researchers have to literally shear genomic DNA molecules into fragments using enzymes (restriction endonucleases), insert these fragments into circular DNA molecules that can be amplified by bacteria (plasmids), and insert plasmids into specialized bacteria strains (transformation). These are subjected to several amplification processes, including one that emits a specific fluorescent signal for each of the

[10] One reviewer noticed that this account might not be compatible with the relational view of data argued for by Leonelli (2016). At first glance, this might be the case: one can say that 'sections of the world' leaving 'marks' by interacting with measuring instruments might imply the idea that data can potentially 'represent' one and one phenomenon only, and hence that data only provide evidence for scientific claims about the specific situation in which they have been generated. But this need not be the case: one can say that data can be potentially used for a wide range of scientific claims well beyond the given circumstances in which they have been generated (as Leonelli does), without denying that the first 'appearance' of data is the result of the interaction between some specific sections of the world and a measuring instrument. In other words, Lowrie's definition does not deny the possibility that data could be subjected to the processes that Leonelli describes in the so-called 'data journeys', and the resulting evidential scope be greatly enlarged as a consequence.

A, C, T, G nucleotides, thus resulting in a sequence of light signals that are detected by a machine, and converted to sequences by an image recognition software. Genomes as biological phenomena are thus (as Rheinberger would say) brought to light by tinkering with biological systems in a way that certain new types of marks/signals/data indicating characteristics of genomes are created. Without material engagement, knowledge cannot be created.

To sum up, there are some epistemic requirements for being an epistemic driver in macromolecular biology. In particular, one has to be able to engage in experimental activities in the way defined above, which means being able to generate new data types (or at least data we know are indications of phenomena) by means of intervening (in the way defined above) materially on experimental systems. Take a fictional, though realistic example. Consider a laboratory where Alice, a postdoctoral wet-lab researcher is carrying out a research project based on an idea that she has discussed with her supervisor. Alice develops a sense of ownership of the project: she studies, prioritizes experimental work, designs individual experiments, and interprets data either on her own or in a discussion with other colleagues, including her supervisor. Her material work and her choices shape the narrative of the project in that they represent a logical and biologically motivated ordering of steps, connected by deductive and inductive activity. She is the one who, beyond taking these steps, is tracing them and choosing a path forward with more or less support and guidance from her supervisor. Alice is, in brief, *driving her project*.

2.2 Consequences for Bioinformatics

There is a sense in which bioinformaticians are not epistemic drivers, which is when theyprovide support for projects of wet-lab biologists. This includes tasks like aligning reads to an annotated genome, performing quality controls, performing hypothesis testing using statistical methods, etc. The computational biologist can merely act as support to help a macromolecular biologist reach their own epistemic goal, e.g. knowing the transcriptional response to the knock-out of a particular transcription factor. This is not something specific to bioinformaticians. Indeed, it is typical of wet-lab biologists as well, for instance when providing orthogonal validation, i.e. an attempt at confirming a result using different molecular techniques or perturbing a system in a different way.

But it is possible to deny, on epistemic grounds, the status of epistemic drivers to bioinformaticians *qua* bioinformaticians. This is reflected in the view that data management and automated data analysis is all that bioinformatics can possibly offer. More explicitly, this can be expressed by saying that bioinformaticians (1) do not do experiments, and (2) do not

have access to the 'material' world of biological phenomena. 1 and 2 are indeed strictly connected: in order to have access to 'biological phenomena', you need to have material access to them, and in order to do this, 'experiments' are required. To put it differently, if the epistemic goals of macromolecular projects (i.e. inferring the configurations of biological phenomena by collecting novel data or generating new types of data) can only be achieved by having a material access to those phenomena, and in order to have this you need to engage in direct experimental activities (in the way defined above), then a bioinformatician is cut out by definition (or at least, under the conception that bioinformatics is only about data management). Because of the lack of material interaction with experimental systems and the inability to do any tinkering (in the way defined in 2.1.1), bioinformaticians cannot generate especially new data types. In other words, bioinformaticians cannot even in principle discover how biological phenomena are constituted, and hence they cannot be epistemic drivers. This epistemic preconception is prior to obstacles related to the social structure of biology that make it hard for bioinformaticians to be epistemic drivers – before even discussing the latter, the epistemic matter has to be addressed.

These considerations are compatible with the evidence gathered by STS studies mentioned in the introduction (e.g. Lewis and Bartlet 2013; Lewis et al 2016). The subordination of bioinformaticians to bench biologists can be summarized by saying that "bioinformaticians do not perform experiments (…) [T]heir practice involves the manipulation of the primary inscriptions produced by biologists, rather than the transformation of the natural world through inscriptions" (2013, p 249). This suggests that materially and directly 'transforming' the natural world through a system of experiments is seen as a necessary condition to be an epistemic driver in macromolecular biology in the first place. The lack of experiments challenges the possibility for computational biologists to be epistemic drivers, and to have their own epistemic goals like wet-lab biologists do.

In summary, there is a view according to which bioinformaticians cannot be, in principle, epistemic drivers, because of their inability to engage *materially* in experimental activities (as defined above) to construct new types of data that can become evidence for answering biological questions.

3. RNA VELOCITY AS AN EXAMPLE OF BIOINFORMATICS EXPERIMENTATION

In the previous section, we have reconstructed the view that bioinformaticians cannot be epistemic drivers. This is based on the assumption that bioinformaticians do not engage in

material experimental activities, and what they do is just data management and superficial analyses. The emphasis on the 'material' and the 'experimental' is motivated by the particular context of this article: macromolecular biology and the prominent role that experimenting has played in it[11]. To counteract this view, we need to rebut both the charge against the lack of 'experimental activities' and against lack of 'material engagement'. In order to argue for these things, we will use a specific case study, namely *RNA velocity* (La Manno et al. 2018). This is an important case for a variety of reasons. First, the computational model of RNA velocity as revealing stable features of a genuine biological phenomenon stems from unique properties of gene expression rather than being the rote application of a model borrowed from other disciplines. Second, the single steps taken by the investigators who created RNA velocity amount to a form of intervention (as defined in Section 2) in an experimental fashion. Third, this experimental intervention happened *in silico*, but traces of materiality can indeed be found, showing that bioinformatics is not as detached from the 'material' as wet-lab biologists seem to think. Fourth, the outputs of RNA velocity go beyond a simple analytical application of statistical models but actually bring forth *a new kind of biological datum*.

This is how we proceed. In 3.1 and 3.2, we illustrate the main aspects of this case study. In 3.3, we introduce some preliminary considerations as to how RNA velocity is a case of bioinformatics experimentation. This is before elaborating a full-blown account of bioinformatics experimentation (Section 4) that is also sensitive to the 'materiality concern' (Section 5).

3.1 What is RNA velocity?

To discuss RNA velocity we need to briefly look at the dynamics of gene expression. For every given transcript in the majority of cell types, the temporal sequence of events (transcription, splicing, modification, export, translation, degradation) is completed in a matter of hours, with transcription and splicing being the longest processes. The different rates at which each of these events happens depend on many biophysical parameters that are both influenced by a cell's current state (e.g., in terms of pH, temperature, concentrations of ions, type and amount of proteins, etc.) and by locus- and transcript-specific features (e.g., sequence, length, subcellular localization, etc.). Greatly simplifying, the relative importance of each of these steps can be classified as follows: if a cell transcribes a gene *G* at a transcription rate that exceeds the

[11] The emphasis on the context is important; in other fields (e.g. physics, astronomy, etc), issues related to experimentation might not cause the same tensions between computational and non-computational scientists as they do in macromolecular biology

degradation rate, then this gene is being up-regulated. If the gene *G* is produced at a rate that matches the degradation rate, it is at a steady state, as its amount does not change in time. Finally, if *G* is degraded at a faster pace than it is produced, it is down-regulated.

Given this picture, some researchers (see Zeisel et al. 2011; Gray et al. 2014; Gaidatzis et al. 2015) had an intuition: if intronic RNA (i.e. the abundance of unspliced, immature RNA) and exonic RNA (i.e. the abundance of mature, already-spliced RNA) could be measured separately for every single transcript that is being made by the cell at a given time, then it is possible to infer how new a transcript is, using the ratio between spliced ("older") and unspliced ("newer") transcript as a proxy for their relative age, and its degradation dynamics. This intuition is based on the knowledge accumulated by macromolecular biology on these phenomena. If the relationship between unspliced and spliced transcripts holds and reveals a temporal trend, it should be possible to 1) model the relationship over time by observing its change across samples taken at different time points and therefore 2) predict the amount of a spliced transcript at a future time point. Given sequencing results at different time points, quantities for spliced and unspliced transcripts can be plugged in a model of gene expression that makes use of ordinary differential equations to describe the relationship between rates of transcription, splicing, and degradation.

RNA velocity is a computational model expressing the relationship over time between unspliced and spliced transcripts. The relation is expressed in such a way that the model can predict the amount of a spliced transcript at a given time. The phenomenon that RNA velocity models is, more precisely, the trajectory of the gene expression state of a cell.

It is important to be more precise on how RNA velocity is related to data, phenomena, and theory, and in which sense RNA velocity creates a new kind of datum. We can understand the relation between RNA velocity, data, and phenomena by considering how these fit into a widely known account, such as Bogen and Woodward's famous view (1988). The phenomenon here is the set of dynamics governing gene expression in a cell, which is a process characterized by stable features that can be identified across different experimental contexts. As a process, gene expression is explained by a number of well-characterized mechanistic models (that, together, constitutes the theory of molecular biology, see Ratti 2020). The trajectory of gene expression is one aspect of the general phenomenon of gene expression. The way this trajectory is represented in the model of RNA velocity is influenced by those mechanistic models. By using RNA velocity, a new type of 'datum' is created, namely data about the trajectory of gene expressions. The idea is that RNA velocity 'models' data on transcripts. One might be tempted to think of this data as 'raw', but data on transcript is nonetheless 'data model', at least in the

sense of "corrected, rectified, regimented, and in many instances idealized version of the data" (Frigg and Hartmann 2020), that we gain from certain experimental procedures. By modeling these 'data models' on transcripts, RNA velocity generates a new kind of data model that provides evidence for a specific aspect of the biological phenomenon of gene expression (that is, its trajectory).

Now that the general framing is clear, let us consider the nature of RNA velocity in depth.

3.2. Discovering through computational tinkering

In order to characterize more precisely the biological phenomenon captured by RNA velocity, bioinformaticians have created a new data type by experimenting computationally, rather than materially. Here we describe the steps of this experiment activity (La Manno et al. 2018).

If we consider a specific unspliced transcript *U*, and follow its maturation at time *t* as the first derivative of *U* in *dt*, it will be characterized by a first order differential equation:

$$\frac{dU}{dt} = \alpha(t) - \beta(t)U(t) \quad (1)$$

where α is the rate at which *U* is transcribed, and β is the rate at which *U* is spliced. As time passes, the total amount of *U* decreases, as unspliced transcripts are continuously being spliced. This means that another measurable quantity, that of spliced transcripts *S*, will increase with time according to another differential equation:

$$\frac{dS}{dt} = \beta(t)U(t) - \gamma(t)S(t) \quad (2)$$

where the rate of production of the spliced transcript *S* is exactly the rate of splicing of the unspliced transcript *U*. Additionally, spliced transcripts are also continuously being degraded at a degradation rate γ; the amount of degraded transcript depends on both the degradation rate and the current amount of spliced transcript.

Assumption 1: the rates are not time-dependent, meaning that for any given time *t*, the values of α, β, γ are constant. Following these assumptions, the first two equations can be rewritten as:

$$\frac{dU}{dt} = \alpha - \beta U(t); \frac{dS}{dt} = \beta U(t) - \gamma S(t) \quad (3)$$

Quantifications undergo total depth normalization, in which each read count for each gene in each cell is divided by the total amount of read counts in the cell. This yields comparable quantities u_i and s_i.

Assumption 2: La Manno and colleagues assume that the splicing rate β is the same for all transcripts, so it is considered to be $\beta = 1$. While this is not exactly true, as splicing can be influenced by several factors, it is a necessary simplification to be able to use the instantaneous measurements (i.e. without temporal information) of *u* and *s*. This also means that all the other rates will be expressed as units of the splicing rate.

The goal of this model is to be able to model the "RNA velocity", expressed as the first derivative of the amount of spliced transcript with respect to time, *dS/dt*. If the model holds, it becomes possible to extrapolate the amount of spliced transcript *S* at a (not too distant) time *t* even if the time is not observed. To be more precise, the model allows us to predict with reasonable levels of confidence the expression dynamics (spliced mRNA) of genes in the near future, thus indicating a direction of change for these genes. The "near future" is limited by the biophysical dynamics of transcription and splicing, i.e. these predictions hold for changes to happen in a few hours.

Taking assumptions 1-2 together, deriving the value of *s* at a given time *t* is achieved by solving the equations for *u* and *s*:

$$u(t) = \alpha(1 - e^{-t}) + u_0 e^{-t} \quad (4)$$

$$s(t) = e^{-t(1+\gamma)} \quad (5)$$

With u_0 and s_0 being the initial unspliced and spliced quantities.

Assumption 3: at steady state there is no change in spliced transcript abundance; mathematically: $ds/dt = 0$.

Taken together, assumptions 2-3 result in $\gamma = u/s$ and $\alpha = u$. If these assumptions were compatible with the complexity and biological features of the phenomena of interest, extrapolating spliced transcript quantifications would be trivial. However, the steady state

assumption (assumption 3) only holds for cells or tissues that do not undergo changes such as differentiation or response to a stimulus. RNA velocity becomes interesting only in a dynamic picture, such as a developmental process; in fact, it would allow researchers to extrapolate "future states" of gene expression based on the data currently available. Moreover, the production rate α is unknown and difficult to measure without specialized experiments. For these reasons, La Manno and colleagues drop assumption 3 and a constant value of α from assumption 1, and make another set of two alternative models, each with its own assumptions.

Model I: velocity is constant, i.e.:

$$s(t) \quad = s_0 + vt \tag{6}$$

This model introduces the velocity parameter as a constant, v , which bears the following relation to the other parameters of the model according to equation (1):

$$v = \gamma s - u \tag{7}$$

Under Model I, and knowing v, extrapolating the amount of spliced transcript at time *t* becomes trivial according to equation (6).

Model II: the amount of unspliced transcripts is constant, i.e.:

$$u(t) = u_0 \tag{8}$$

meaning that equation (2) can be reduced to a much simpler form:

$$\frac{ds}{dt} = u_0 - \gamma s(t) \tag{9}$$

whose solution for *s(t)* is also simpler:

$$s(t) = s_0 e^{-\gamma t} + \frac{u_0}{\gamma}(1 - e^{-\gamma t}) \tag{10}$$

Both models I and II effectively require the development of a computational procedure to estimate the gene-specific degradation parameter γ to extrapolate *s(t)* and determine gene-specific velocity values.

At a steady state, i.e. working under assumption 2, thus setting the velocity to 0, it is possible to estimate γ as setting $v = 0$ in equation (7) gives

$$u = \gamma s \tag{11}$$

meaning γ can be found using a simple linear regression, using the quantities of *u* and *s* across different samples or cells. However, as explained earlier, assumption 2 only holds under specific biological conditions.

The dynamics of each specific transcript can be represented as the progression of the combination of spliced and unspliced quantities, i.e. different solutions to a system of differential equations. The geometrical representation of these solutions constitutes a phase portrait (Figure 1). More precisely, for every acceptable pair of u_i and s_i values - that is, for every pair that can represent a solution to these equations - there is a point in space; connecting these points along their variation in time creates the phase portrait. This representation is useful to understand the relationship between the parameters, the quantities, and their progression in time. According to equation (11), equilibrium points (i.e. points where velocity is 0) are reached where γ is equal to *u/s*, or where both *u* and *s* are 0, or when $\alpha = u$. Then, the fit of a regression line going through the diagonal of the phase portrait represents the steady state approximation for γ (Figure 1A); in other words, a simple linear regression coefficient will be accurate if and only if all samples are at the equilibrium points of the phase portrait. However, as discussed previously, samples/cells undergoing differentiation or responding dynamically to a stimulus will be populating many other parts of the phase portrait, making a linear fit on their *s* and *u* values severely biased.

Accordingly, the RNA velocity authors use an "extreme quantile fit": rather than trying to calculate the coefficient using all points in the phase portrait, they only consider points that lie at the extreme of their distribution (Figure 1B), thus getting closer to the steady state assuming degradation rates do not change along the trajectory.

This procedure, however, only works well when the extreme quantiles are close to the steady state. There may be genes that are up-regulated late or down-regulated early, so that we do not observe their steady state in the time period sampled in the experiment; in this case, their extreme quantiles will lie in the middle of the phase portrait, meaning the fitted γ will be still biased (Figure 1C, D). For these cases, the authors developed yet another model termed "structural fit", which accounts for the number of exons, length of introns, and number of internal priming sites that can be captured by the sequencing technology.

Building these models require an understanding of the unique properties of transcripts and of the sequencing procedure. Thus we can see how so far the development of RNA velocity as an approach requires a high degree of *data analytics* (data processing, transformation, modelling with different mathematical approaches) but also of *software development* (coding all these implementations in an efficient and usable way for other practitioners): La Manno et al. had to write code to perform data preprocessing, phase portrait estimation, model fitting, estimation of the velocity vectors and their visualization, in a programming interface that can be applied to commonly used data representations for single cell RNA-seq.

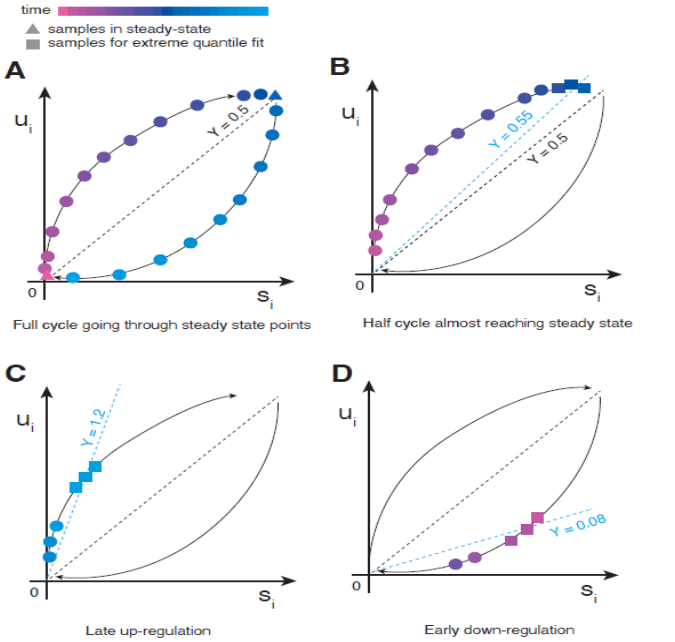

Figure 1. Phase portraits for RNA velocity. Blue lines show fitted coefficients. Adapted from La Manno et al. 2018.

Indeed, it can be useful to pause at this point to reflect on how the calculations undertaken under different assumptions, resulting in two alternative models, and the different attempts at fitting linear models to derive degradation rates amount to experimental tinkering. While the intuition of modeling gene expression using simple differential equations is far from new, there are a few important novel aspects in this approach.

The first is the concept of RNA velocity itself, which finds an important application in the field of single cell transcriptomics - an important intuition by Sten Linnarsson and Peter Kharchenko. As it is widely known, the innovation of single-cell transcriptomics lies in its ability to capture a large number of individual cells within a tissue/organ, as opposed to the "bulk" sequencing of the transcriptome of a tissue/organ. Therefore, analyzing a population of cells that is undergoing a transition in an un-synchronized fashion, such as a developmental process, means collecting a snapshot comprising different phases of the process itself, within a

reasonable time frame. Conversely, in a "bulk" setting where the transcriptome of each cell is mixed, there is only one such average phase per sample. It becomes evident that population-level temporal dynamics (often called "pseudo-temporal") can be inferred in a single cell dataset by virtue of these cells being, individually, at different stages of their progression along a specific biological trajectory – which is described by RNA velocity.

The second important innovation by the authors of the original RNA velocity paper lies in the techniques for the visualization of velocity. The opportunity to derive cell-specific velocity vectors within a dataset at single cell resolution brings forth an additional layer of complexity to the canonical single cell data analysis outputs. Originally, the visualization of single cell data posed an important challenge: if each cell is embedded in a space according to its gene expression values, meaning every cell is represented by a point, the coordinates of this point will be determined by the numeric expression values of *n* genes: points will exist in an *n*-dimensional space. For this reason, dimensionality reduction techniques have been leveraged to reduce complexity while retaining meaningful relationships in a two- or three-dimensional representation: in other words, points (cells) that are close together in this visual space are supposed to be similar, while points that are far away are supposed to be different. Several techniques have been proposed, at different levels of granularity: t-stochastic neighbor embedding (t-SNE, van der Maaten 2008), uniform manifold approximation projection (UMAP, McInnes 2018), partition-based graph abstraction (PAGA, Wolf 2019), similarity weighted nonnegative embedding (SWNE, Wu 2018), diffusion pseudotime (Haghverdi 2016), to name a few. And, as researchers routinely discover, visualization techniques may be biased, imprecise, or contain assumptions that are at odds with what we know about the biological systems they are meant to represent, giving way to new, improved visualizations that should be "more faithful" to the underlying biology. Visualization of high dimensionality data is an active field of experimentation in computational biology (and machine learning in general) and, as every experimental field, it proposes partial solutions with advantages and pitfalls (see Chari and Pachter 2023 for the case of UMAP). These representations play a central role in the analysis of single cell data, as they are not only ways of summarizing an analysis output, but they are also *de facto* data models that are used for discovery, inference, and validation of hypotheses – a point emphasized by Stevens in his ethnography of bioinformatics (2013). One of the most important outputs of the RNA velocity procedure can be considered an enhancement to these visualizations: a two-dimensional representation of the velocity vectors, pointing to the future state of single cells, within the "transcriptional space". The authors of RNA velocity devised a technique to draw velocity arrows on top of two-dimensional

visualizations that were previously created, either at the level of single cells, or as a "vector field" that shows a summary of local velocity at every point of the transcriptional space. Thus, by looking at a UMAP visualization of single cells and overlaying their velocity vector field, researchers can literally see whether a certain cell population is progressing towards another, thus inferring that a differentiation progress is taking place with a certain directionality and intensity.

3.3 RNA Velocity and Experimental Activities

What we think RNA velocity shows is that bioinformaticians engage in experimental activities, understood as investigative activities involving interventions that, by manipulating existing data, can even create new data types, exactly like traditional macromolecular biologists.

The estimation of RNA velocity vectors is non-trivial, and presents many challenges to the original authors of the method. They make use of several models, alternative ways to fit degradation coefficients, and simulations that test the extent to which their models hold given differences in gene expression levels, equation rates, and their temporal dynamics. There was no *a priori* guarantee that RNA velocity would represent a relevant biological phenomenon once single cell transcriptomic data was used and processed. Comparisons to real-world datasets with different levels of ground truth are included as a validation of their experimental procedure.

Taking these considerations together, it can be argued that 1) the quantification of transcriptional dynamics in single cell data does not require *ad hoc* experimental procedures, rather a repurposing of existing data; 2) the extrapolation of a cell's future transcriptional state is not only a biophysically motivated ordering of cells along a trajectory, but also a measurement of an unobserved instantiation of such a trajectory; 3) extensive tinkering with different models and assumptions was required to arrive at a final, usable data model. But RNA velocity is being investigated also by other groups, in direct competition with the original picture. As it happens for the discovery and characterization of other new biological phenomena, the publication of the RNA velocity paper sparked many enthusiastic reactions, and the community quickly started building on top of the original models and results. In fact, several alternative versions of RNA velocity estimation were published, which made use of different assumptions, different models, different representations and had different software implementations. The corpus of experimental work on the *field* of RNA velocity is growing – estimating RNA velocity, as an experimental activity, has a life of its own.

A large and complex study (Gorin et al., 2022) published a few years after the first RNA velocity paper aims at laying down more rigorous foundations for the method, which implies an inevitable critique of the original work and most of its derivatives. We will not go through the details of the study as it is a very exhaustive treatise of the biophysical foundations of the model, but we want to highlight what we perceive to be important contributions, both to the field *per se* and to our argument in particular: the RNA-velocity strategy generates *a new type of biological datum that is evidence for a specific biological phenomenon worthy of studying in its own right*. The Gorin et al. study highlights the great potential of RNA velocity approach(es), but at the same time sheds light on several issues, motivated by a computational experiment: the same dataset analyzed through two different RNA velocity implementations yields two very qualitatively different results (Soneson et al. 2021; Gorin et al. 2022). The first issue is the definition of RNA velocity itself, which can be interpreted in seven different ways. The second issue concerns different processing pipelines which potentially render some of the assumptions invalid, in particular considering spliced and unspliced molecules two mutually exclusive species and thus over-simplifying the complexity of alternative splicing. Then, assumptions made by different implementations are also quite diverse. Additionally, they critique the visualization of RNA velocity itself, following previous work by the same authors in which they address the larger issue of whether a visualization through severe dimensionality reduction is properly representing a biological phenomenon or not. (Chari and Pachter, 2023) Chari and Pachter go through a rigorous study of these assumptions performing other computational experiments (such as the application of RNA velocity estimation to a dataset with no differentiation or stimulus) and conclude that the current implementations of RNA velocity reduce the complexity of the quantities they are trying to model, are lacking in biophysically motivated foundations, require restrictive assumptions, make use of arbitrary parameters and as a consequence do not result in reliable estimations of a future cell state; their critique of current implementations goes as far as questioning whether RNA velocity can be useful at all or whether something can be salvaged by asking the Biblical question: "*is there no balm in Gilead?"*. In the last few years, the Pachter group has worked on more biophysically motivated models of transcriptional activity which show, by the application of different models and mathematical frameworks, how genes can be classified in different ways (Gorin et al. 2022), and how a precise modelling of stochasticity in gene expression and its measurement is required to describe transcription mechanistically using single cell sequencing data (Gorin et al. 2023). Interestingly, in this article Gorin and colleagues explicitly mention tinkering with

their virtual system by way of "manipulation of generating functions" (Gorin 2023) in purely experimental ways.

To summarize, RNA velocity has been developed through computational experimentation, made available as an analysis tool, been experimented with and heavily refined (with some refutation of its assumptions and models) as a result of both additional experimentation and mathematical formalization, and has taken on a life of its own. We argue thus that through RNA velocity *a new type of biological datum* is constructed that does not stem from modifications in wet-lab experimental procedures, but rather from an elegant and complex *in silico* system of experiment. By creating a new type of biological datum that can provide evidence for a specific biological phenomenon (i.e. the trajectory of gene expression states), the computational work seems to achieve the same kind of result that material tinkering performed by wet-lab biologists can achieve.

4. AN EXPERIMENTAL ACCOUNT OF BIOINFORMATICS

In the previous section, we have reconstructed an example of bioinformatics practice where a new data type providing evidence for a specific biological phenomenon is created through various experimental activities done *in silico*. In this section, we describe these activities at a more general level, by constructing a comprehensive account of the dimensions of bioinformatics as an experimental discipline consisting of three dimensions (data management; analytics, development). This account builds on previous analyses and observations of bioinformatics, most notably (Stevens 2013; Leonelli 2016; Strasser 2017) and discussed in Appendix 1. The facets of our account will be illustrated by referring back to the example of RNA velocity.

4.1 Bioinformatics: a Tripartite Account

Our account of bioinformatics counts three dimensions.

First, there is *data management*. As mentioned earlier, this is a central aspect of bioinformatics practice, given the importance of databases. It includes those practices geared at creating, maintaining, interfacing with and creating connections among biological databases in virtual spaces. Data management thus consists of creating standard formats, an easily accessible and navigable infrastructure, secure storage and updated records; from an end user perspective, it is the management of laboratory archives with special regard for high throughput data, making sure that datasets are properly stored and shared together with their metadata, and

ensuring the reproducibility of the raw data processing steps[12]. In the case of RNA velocity, the practices associated to the management of single cell transcriptomics (which have been built on the foundations of the transcriptomics data management ecosystem) are instrumental in creating reproducible analyses, such as the storage and distribution of spliced and un-spliced count matrices.

A second dimension of bioinformatics is *analytics*, which is the application of more or less established statistical models and computational procedures to gain a first level of biological interpretation of a given experimental outcome. Analytics include processing raw data into quantities of features of interests (such as genes, proteins, chromatin regions, etc.); applying quality control procedures to distinguish signal from noise, rule out technical artifacts, and remove systematic biases; applying mathematical frameworks to identify patterns and score relevant differences between experimental conditions, together with a measure of their uncertainty; visualizing results in a clear and informative way; etc. These aspects have been also emphasized by Stevens (2013) in his analysis of the epistemic roles of data visualization, and more recently by Leonelli (2019) in her ethnography of the SureRoot project. What these – and other examples – show is that modern analytics consists of different steps that can be combined together in different and novel ways; far from being 'automated', analytics allows the creation of analytical pipelines with varying degrees of flexibility. In the case of RNA velocity, analytics is a critical aspect, as it is a specific combination of several processing and mathematical modeling steps that creates a new data type, and has become after its development a rather standardized step in several single cell transcriptomics analysis pipelines.

Finally, *development* consists in the invention and programming of new mathematical and statistical frameworks, or the optimization of previously available frameworks, with a specific type of biological question in mind. A bioinformatician involved in development aims at writing software that tackles extant challenges in the generation and interpretation of results, such as identifying and implementing the use of the correct statistical distribution for a certain type of datum; integrating different data modalities to enhance the discovery of biologically relevant phenomena; using "first level" analytical results to predict more complex behaviors of a biological system, or non-trivial ways in which this system can be modified, and so forth. Writing software not only entails the theoretical exercise of finding the most appropriate models, operations or representations for the data, but also the practical aspect: implementing

[12] The data management dimension of bioinformatics has received significant attention both from a historical point of view (see in particular Stevens 2013; Strasser 2017) as well as from the point of view of the creation of new scientific roles such as data curators (Leonelli 2016)

the algorithms, optimizing them, creating a usable interface and its documentation. A few examples of these new frameworks and implementations will be discussed in section 4.2. There is empirical evidence suggesting that development is experiencing an exponential growth: even just considering the short time frame between 2017 and the first quarter of 2024, and limiting ourselves to the field of single cell biology, the number of bioinformatics tools has quickly surpassed 1700 units (Figure 2A, Zappia 2018). Similarly, when querying the number of R packages distributed from the Bioconductor project's first release in 2002, we observe a similar trend (Figure 2B). In the case of RNA velocity, the implementation and optimization of the processing and modeling steps, together with the creation of a user-friendly software interface (and several other improvements and iterations from other computational biology groups) that we have described in the previous section is a classic example of development..

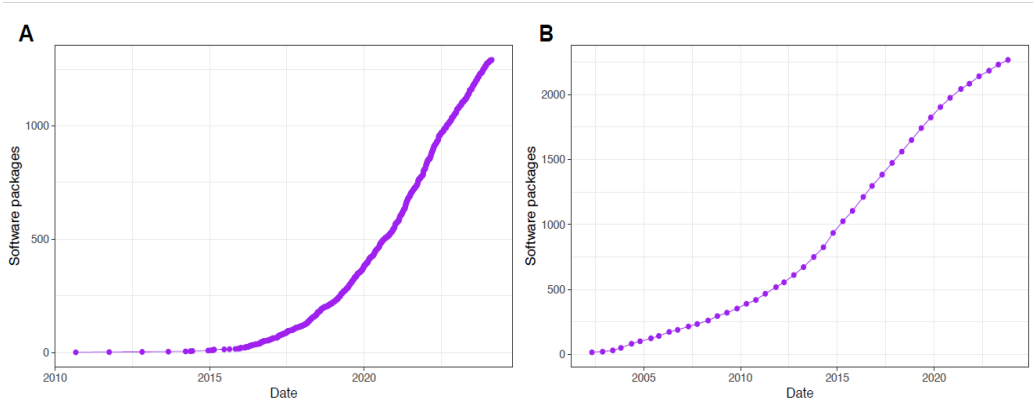

Figure 2. Exponential growth of number of software packages over time. A: single cell analysis tools, regardless of programming language. Source: https:/ /www.scrna-tools.org/, captured in March 2024; B: packages released on Bioconductor, regardless of the application. Source: https:// bioconductor.org/about/release-announcements/, captured in March 2024.

These three aspects (i.e. management; analytics; development) are fundamental integrated ingredients of bioinformatics as a discipline. As such, they often complement each other and coexist within many declinations of bioinformatics practice, as Strasser, Leonelli, and Stevens have noted, even though without using the terminology employed here.

For instance, management can imply development, as bioinformaticians who want to distribute their software to a large community should, at a minimum, provide thoroughly tested software that has as few problems as possible, make it easily accessible and easily findable through metadata, provide clear documentation and instructions on how to use it, and - if the software is released as open source, which is in most cases - store the source code in repositories that allow version control. In order to make this collective endeavor easier and standardized, bioinformatics developers started projects such as Bioconductor (Gentleman 2004; Huber

2015), which hosts more than 1250 packages for biological data analysis and is maintained by the community on a volunteer basis[13].

Analytics and data management are intertwined as well, especially regarding the reproducibility of data analysis. Whereas wet-lab experimental procedures are usually succinctly described in the methods section of a paper, leaving their complete description to protocols (commercially or academically published), the gold standard for analytics reporting is to provide other researchers with the exact steps, i.e. the code used for the analysis, together with the expected outputs and the necessary inputs. This, in turn, requires bioinformaticians to provide their colleagues with access to the same software they used, as some outputs could depend heavily on the software version that was used. Thus, analysts need to manage representations of their workflows - aptly named "notebooks" - and digital snapshots of the software they used - "images" or "containers" - to ensure that their results are reproducible by anyone with sufficient skills.

But the most interesting integration is the one between analytics and development. On the one hand, developers usually master some facets of the analysis toolkit, even just to be able to benchmark the results of their newest algorithm against gold standard applications, or to generate data representations upstream or downstream of their inventions. On the other hand, analysts can combine different tools crafting pipelines which, at some levels of complexity, can be considered akin to development. In fact, analysis tools are often developed with the Unix philosophy in mind: programs should do one thing, and do it well (McIlroy 1978). This translates to a highly modular analysis workflow in which every step can be carried out by several alternative approaches and/or tools. An expert analyst combines these tools and, in many instances, refines their input writing code that can be in the same language as the one of the tools they use. In some cases, entire analysis workflows can be packaged as single one-stop solutions, showing how blurry the line between analytics and development can be.

4.2 Soft and hard experiments in bioinformatics

We have mentioned throughout this article that bioinformatics should be considered an experimental science, as much as macromolecular biology. But how exactly? We claim that the integration of analytics and development plays a central role.

[13] Bioconductor provides an infrastructure for storing, distributing, updating and checking the integrity of a wealth of bioinformatics software.

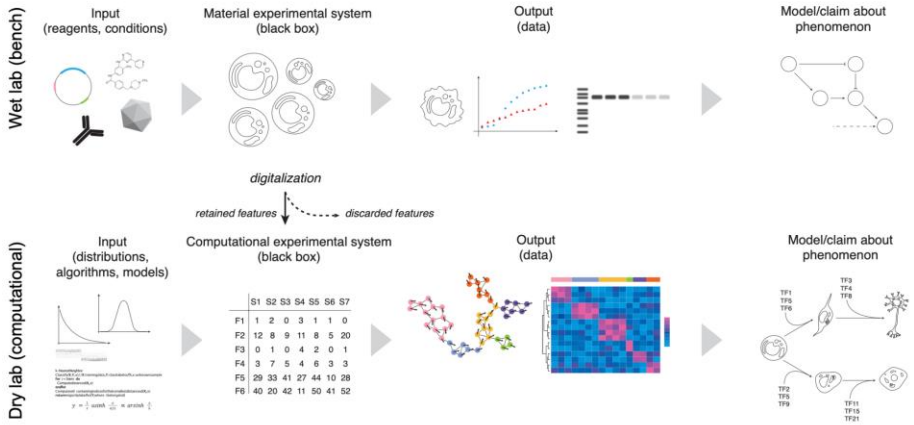

Figure 3: schematic representation of the parallel between material (wet lab) experiments and computational experiments.

First, let us draw a parallel between the 'wet-lab' biology pipeline and the computational one (Figure 3). In the case of macromolecular biology, the starting point is a question about how biological phenomena are produced and/or maintained, as these are often opaque. A material experimental system (which is taken to materially embed the biological phenomenon a biologist wants to explain) is perturbed or tinkered with certain inputs (e.g. reagents, conditions). The results of this tinkering are output data that are taken to be evidence for certain claims or mechanistic models (Craver and Darden 2013) about the biological phenomenon under scrutiny. As noticed in Section 2, this is not mindless tinkering. In fact, conditions of calibration, robustness, internal/external consistency, and specific biological questions will constraint the tinkering. In the case of bioinformatics, an analogous dynamic is at play. Consider the case of RNA velocity. The material experimental system is converted into a virtual system - by digitizing some of its features and embedding them in an appropriate numerical representation - but *it is tinkered with anyway*. This is not necessarily a specific aspect of virtualisation. In fact, any experiment consists in isolating specific, robust signals from a (biological) system and recording them through an apparatus. However, recording a high amount of observations in parallel and embedding them into a virtual system - what biologists call 'high throughput' technologies - allows to retain several 'hidden' relationships between data points or features for which we do not necessarily know the data generating process. In this virtual system it is possible to do new work that is experimental - though not material - to identify new relationships, new signals, new secondary inscriptions through new techniques. These novel biological insights are particularly interesting because they retain a 'semi-material' property given the material origin of the virtual system, and their reliability is

almost always confirmed by comparing them to a material experiment. Bioinformaticians interact with such a system by modifying inputs and/or rules, and monitoring changes in outputs. The way bioinformaticians interact with such systems makes ample use of the 'confidence-building' strategies exemplifying good experimental activities in material systems as emphasized in Section 2, including calibration, robustness analysis, and checking for consistency. Following these strategies reflects the notion of experimenting that we have formulated previously: bioinformatics is not just tinkering; rather, it involves controlled settings (which allow monitoring), biological hypotheses entertained, and recorded unforeseen consequences and effects. For instance, bioinformaticians do not know *a priori* what kind of statistical distribution or differential equation best fits the description of a particular biological phenomenon; indeed, they always clarify certain assumptions that make their models tractable, and test a set of proposed models with respect to "ground truth" data. However, ground truth is not always, if at all, attainable in biology, forcing researchers to use reasonable approximations or resort to orthogonal validation. Thus, bioinformaticians deal with data models that have potentially surprising behaviors due not only to their non-deterministic nature, but also to unobservable variables. As an example linked to the case of RNA velocity, consider the cellular composition of a tissue as inferred through single cell RNA sequencing and unsupervised clustering. Most currently utilized clustering algorithms do not require the user to specify how many clusters are expected, nor their size, or their degree of separation. A clustering algorithm that operates in ideal conditions should be able to partition the data in biologically relevant (and, possibly, experimentally separable through other means) units such as "cell types" or "cell states". A surprising behavior thus would be the identification of a previously unobserved (and unexpected) intermediate state between two well characterized differentiated cell types, or the existence of a continuum bridging what were previously thought to be isolated, highly stable transcriptional profiles. More precisely, what is surprising here is the fact that a specific novel clustering algorithm may or may not reveal "new" biologically plausible aspects of the virtual system that other algorithms could not reveal before. Tinkering on the data with different clustering techniques, or inventing an entirely new clustering technique, is what we refer to as experimental in this setting. Several approaches can be used to estimate or infer the unobservable variables of interest, which can then be correlated with the biological nature (e.g. whether these variables overlap with a disease or mutant state compared to a healthy/wild type control). Changes in algorithms, models, parameters all amount to different experimental procedures on the same data model, resulting in different outcomes. The evolution of bioinformatics tools - through analytics and development - shows

that indeed our understanding of the same data can be furthered by testing and refining procedures. For instance, in the context of RNA-sequencing and differential expression analysis several authors over the years have suggested different statistical distributions and models to deal with read count data (Marioni et al. 2008; Anders and Huber 2010; Robinson et al. 2010; Trapnell et al. 2010; Law et al. 2014): using t-tests, linear models, generalized linear models for Poisson-distributed data, Negative Binomial GLM, etc. Eventually, after several experiments and benchmarking studies (Robles et al. 2012; Soneson et al. 2013; Germain et al. 2016), the field appears to largely favor the use of linear models and/or negative binomial generalized linear models with variance shrinkage, although it has been argued that other methods are more precise and reliable in particular settings (Li et al 2022). This type of tinkering is done by integrating and modifying analytical tools and developing new software or computational infrastructures that can host the right set of tools. These are all forms of experimental activities.

We can be even more precise and distinguish between *soft* and *hard* forms of experimentation. Soft experiments in bioinformatics are the attempts at using new (or old, but refined) approaches and techniques to gain deeper understanding of data: existing approaches are being repurposed or extended to deal with biological data beyond their original scope, with no *a priori* guarantees regarding their reliability, robustness, fidelity to the natural process or interpretability. It is tinkering in a controlled setting by recording unforeseen consequences. Bioinformaticians operating on data models by applying analytical steps that create new representations of the model and new data types is the *hard experimental nature* of bioinformatics[14]. In the soft experimental framework, the novelty lies in the *use* of the tool, and not in the tool itself. Conversely, in the hard experimental framework, the novelty *lies in the tool itself* and in *the new type of data* generated that can constitute indications for claims about biological phenomena. Regardless of whether an approach is soft or hard, it should be considered experimental by virtue of the tinkering, possibility of unexpected results, internal cohesiveness of the digitalized experimental system, and biological focus.

There are several examples in the history of bioinformatics that can be described using our framework of hard and soft experimentation. Take for instance the creation of differential expression tools such as DESeq2 (Love et al 2014). This is a soft experimental activity: the use

[14] Please note that saying that these bioinformatics activities create 'models' does not exclude that the activities are 'experimental'. To paraphrase Parker (2009), models are types of representation, while experiments are investigative activities involving intervention. As such, there is an experimental side of modeling (Peschard and Van Fraassen 2018), and it should not be very surprising.

of generalized linear models and Bayesian variance shrinkage greatly predates RNA sequencing, but its application and successful implementation required tinkering and innovation. The resulting data, i.e. fold changes (effect sizes) and corresponding statistical significance values, do not belong to novel data types as they are basically the same type of result one would construct from other tests in which group means are compared, with or without computational tools (e.g. in the case of qPCR). This is why we can consider the invention of these tools as a case of soft experimentation. The case of Gene Set Enrichment Analysis (Subramanian et al. 2005), instead, constitutes a hard experiment (as it is the case of RNA velocity): ranked gene lists were tinkered with in ways that created a new data type, the Enrichment Score, a numeric value whose sign and magnitude is indicative of the activity of a pathway in the comparison of global gene expression programs across conditions. The Enrichment Score is also constructed by comparing observed data to an empirical null distribution built by random permutation of rankings, a common approach in statistical testing that is akin to constructing a virtual negative control.

Enumerating which bioinformatics tools constitute hard or soft experimental approaches is beyond the scope of this article, but we supply a table with a small subset of examples spanning the domain of transcriptomics and other high-dimensional genomics data analysis tasks.

Tool/approach	Reference	What it does	Type	Resulting data type (new?)
DESeq2	Love et al. 2014	Differential gene expression analysis for RNA sequencing using generalized linear models	Soft	log2(fold change), p-value, Wald statistic and associated error
limma	Ritchie et al. 2015	Differential gene expression analysis for microarrays using linear models	Soft	log2(fold change), p-value
sequence-based phylogenetic trees	Fitch and Margolias 1963	Use of genomic and/or protein sequence alignments across species to draw a phylogenetic tree	Soft	phylogenetic tree, inter-species distances
graph-based clustering	Blondel et al. 2008, Traag et al. 2009, Xu and Su 2015	Identification of communities/clusters of cells in an undirected graph built in a low-dimensional transcriptional space	Soft	cell type/cell state clusters
random forest regression	Díaz-Uriarte and Alvarez de Andrés, 2008, Huynh-Thu 2010	Identification of a small number of genes to classify samples; regulatory networks	Soft	gene sets and their classification power; regulons
GSEA	Subramania n et al. 2005	Quantification of the regulation of a pathway in transcriptomics datasets	Hard	Enrichment Score (new)
RNA Velocity	La Manno 2018	Prediction of future transcriptional states of single cells based on their splicing dynamics	Hard	RNA velocity vectors and vector field (new)
Trajectory inference	Trapnell et al. 2014, Haghverdi et al. 2016	Distance- or similarity-based ordering of cells along transcriptional continua	Hard	Pseudotemporal ordering and trajectories (new)
NovoSparc	Nitzan et al. 2019	Tissue-level patterning prediction from low dimensional embedding of single cells through optimal transport	Hard	Gene expression cartography (new)
CellOracle	Kamimoto et al. 2023	Prediction of shift in transcriptional dynamics following a virtual KO using RNA velocity and chromatin accessibility	Hard	In silico knock-out (new)

Table 1. Examples of soft and hard experimental approaches in bioinformatics

5 WHAT ABOUT MATERIALITY?

By conjuring the possibility of hard and soft experiments in bioinformatics, we want to argue that bioinformaticians can indeed be epistemic drivers, because they generate new data or even new types of data that can potentially constitute new biological knowledge, and they do this by experimenting in an analogous way to how macromolecular biologists do.

However, remember that the argument against bioinformaticians being epistemic drivers was not only about experiments; 'materiality' was also involved. Maybe bioinformaticians do experiments; but given that they do not materially manipulate and generate data (or, to use Lewis and Bartlett's conceptual apparatus, they do not generate *primary inscriptions*), then they cannot be epistemic drivers. By relying again on the case study of RNA velocity, our response to the 'materiality' concern is twofold. First, we show that the importance assigned to 'materiality' is misleading. Second, even if materiality was indeed that

important, there is still space for something that we call *semi-materiality*, which basically applies to what biologists do, at all levels, be they wet-lab biologists or *in silico* biologists.

Let us start by showing how misleading the idea of 'materiality' can be. Intuitively, materiality is deemed important because it provides a 'more direct' access to biological phenomena. Given that bioinformatics lack this 'direct access', then they do not have the same grasp of biological phenomena that wet-lab biologists have. However, it is just not the case that it is in virtue of materiality that we have a more or less mediated access. In fact, the case of RNA velocity shows that, while we need a preliminary 'material origin', virtualizing the data (or making them 'semi-material', as explained below) is what provides us a better access to the phenomenon itself – just with traditional material access, the phenomenon captured by RNA velocity is inaccessible. Moreover, consider the variety of experimental systems that biologists use: *in vivo, in vitro,* animal models, etc; these are often only proxies for various biological phenomena, and hence one may say that the access to phenomena is nonetheless mediated, and materiality plays no substantial role in making these systems more inferentially reliable.

But let's say that materiality is indeed important (even if the 'importance' is vague). How should we address this? Our response is that we are not advocating for an exclusively *in-silico* knowledge generation process: *the origin of the data is always material*, unlike in some cases of computer simulations. This aspect is not appreciated enough. We can draw a parallel to so-called 'virtually, experiments', namely nonmaterial experiments on semi-material objects. In (2003), Morgan describes computational experimental activities to investigate the strength of bones. Given the challenges of assessing strength in 'material' settings, one line of investigation was to convert a real cow hipbone into a computerized image – cutting bones into thin slices, taking specific pictures of them, re-assemble these in high-quality 3-d computerized images, and then intervening on them by means of various models. Morgan emphasizes how this process "retains a high degree of verisimilitude of structure for each particular bone sample" (p 223). By conserving important structural features of bones in the process of recording and converting, those computerized images have a 'semi-material status': intervening on the 3-d images is *de facto* an experimental activity, where mathematical models are used *as* experimental instruments. Virtual experimental systems are akin to *semi-material objects*: biological features are recorded, then converted, but nonetheless conserved. In the case of RNA velocity, we have seen that certain 'physical' aspects of data sets are conserved in the virtual experimental systems, such as the differences in spliced and un-spliced transcript abundances in single cells.

The material origin of data has other consequences too. Given that 'aspects' of materiality are conserved, there is the risk of importing factors that may be confounding – , to paraphrase Morgan, bioinformaticians have to consider "all conditions and factors that are likely to interfere with the process of interest" (Morgan 2003, p 219), exactly as traditional experimenters. In the case of RNA velocity, these confounding factors are the stochastic aspects of transcript quantification given by the material process of transcript capture and reverse transcription of a very low input; the presence of dying/stressed cells whose gene expression does not represent a physiologically relevant cellular state; the lack of complete knowledge of the structure of transcripts and their spliced forms. Moreover, the material origin comes with the possibility of 'discovering' something hidden in the data that, for various reasons, could not be separated materially. This is noteworthy in contemporary biology: in cases like high-throughput recordings, an impressive amount of observations is recorded, and these observations have hidden relations for which data generation processes are unknown. 'Virtualizing' such high-throughput experimental systems means *recording* (by means of conversion) these 'hidden relations' on a virtual experimental system, and *transforming* signals to create new data or new data types, as we have shown. Tinkering with multi-dimensional data within a controlled virtual experimental system means being able to separate signals that, in the normal material laboratory setup, would just be impossible to distinguish. As much as in normal laboratory conditions results are 'produced' by intervening on a (material) system, here results (e.g. new data or new data types) are produced by intervening on the (virtual/semi-material) experimental systems, unlike typical cases of modeling where one derive results just by means of mathematics (Morgan 2003). And it is in virtue of the fact that something is conserved in the transition from 'material' to 'virtual' that inferences based on computational experimental activities can be, at least in principle, reliable. Of course, this does not mean that reliability is established purely *in silico* – in fact, orthogonal and functional validation from wet-lab biologists is still required (but this is true of any paradigm, even in the wet one).

But one can push the 'materiality' argument further, in somewhat unreasonable ways. One way to do this is by appealing to the intuitive distinction between 'primary' and 'secondary inscriptions' (Lewis and Bartlett 2013). The idea is that bioinformatics data might have the 'semi-material' dimension we have argued for, but they are still 'secondary inscriptions' - by not being able to generate 'primary inscriptions', bioinformaticians cannot in principle meet the epistemic desiderata for being drivers. This is a strong claim, likely to end any discussion. However, the distinction between primary and secondary inscriptions (Lewis and Bartlett 2013) is misleading. For instance, in sequencing a sample, which 'data' is considered a primary

inscription? Is the sample itself? But the 'sample' *per se* does not constitute a datum, and in order to become data (e.g. a sequencing read) is manipulated by various technicians, including computational technicians (Stevens 2013). Therefore, the primary inscription is a sample manipulated to become sequencing data. This shows that primary inscriptions are both material and *in silico* at the same time, and that they are co-produced both by traditional biologists and bioinformaticians. This means that there is not really any substance to the 'primary vs secondary' distinction: all data is likely to be semi-material. But if this is the case, then in principle there is no difference between the data used by bioinformaticians, and data used by wet-lab biologists, at least from this perspective. In conclusion, this shows that even the second concern (the materiality concern) has not any robust substance, and hence there is in principle no reason why bioinformaticians cannot be epistemic drivers.

6. CONCLUSION

The fundamental role played by computational biology in most life sciences projects has grown at a quick pace, so much that international consortia such as the Human Cell Atlas (Regev et al 2017) require an effort in coordinating, developing, testing and communicating computational methods for data storage, analysis and visualization that is far beyond the - still impressive - work required to generate all the single cell atlases. In increasingly more cases computational biologists use the generation of an atlas as a good testing ground for a new computational method (e.g. Stephenson et al 2021), or they drive highly complex analysis efforts to establish best practices in the field with no additional data generation required (e.g. Luecken et al 2021). These computational scientists do control the narrative of their projects and are fully equipped by their environment to be *epistemic drivers*: this is, we claim, bioinformatics as a proper discipline, rather than just support for wet-lab biologists. This article is only a first step towards a comprehensive characterization - both philosophical, historical, and institutional - of bioinformatics.

To conclude this piece and introduce future works, we formulate in the remaining space one open question that we did not address in depth for the sake of brevity, but still deserve a mention in our conclusions, and further elaboration in its own merit.

We have motivated the need for a complete account of bioinformatics practice by mentioning the problem of epistemic alienation. We have co-opted the term alienation directly from Karl Marx's posthumous *Economic and Philosophic Manuscripts of 1844*, where alienation (*Entfremdung*) is defined in terms of the estranged relationship between (1) the laborer and the act of production, (2) the laborer and the product itself, and (3) the laborer and

their very own essence (*Gattungswesen*). Often, there are power dynamics in the biological community preventing bioinformaticians from having access to important decisions regarding experimental design, hypotheses to be tested, and raw data generation procedures, (1); the bioinformatician thus receives data that they are supposed to analyze and convert into biological knowledge, with little room for original interpretation and contribution to the narrative of the study, or freedom to suggest additional experiments (2); thus, the bioinformatician is systematically denied the status of epistemic driver, although they still consider themselves (and expect to be considered) scientists (3). Our intuition is that the same Marxist lens allows us to look at the opposite face of the coin as well: a wet-lab scientist who produces high throughput data is often unable to follow it up through its analysis, leaving important choices in the hands of bioinformaticians (1), who are still required to translate the wet-lab's scientist experiment into biological knowledge (2); without this intermediation, the wet-lab scientist cannot carry out their project and control its narrative, a fundamental aspect of epistemic drivers (3). The open question regards the standing of our theory in the real world: does recognizing the experimental nature of bioinformatics provide a cogent and natural justification for bioinformaticians to become epistemic drivers? Are some bioinformaticians more experimental than others, and does this correlate with their ability to become epistemic drivers? And, if a transformation could be brought upon the field by shifting norms and practices, would the epistemic alienation experienced by both wet-lab and computational scientists be greatly reduced, if not entirely dissolved, creating more collaborative and harmonious research environments?

Acknowledgement: We would like to thank participants of the Linz-Wien work in progress group, Pierre-Luc Germain, Hallam Stevens, Sabina Leonelli, and Sarah Langley for valuable feedback on very early draft of this manuscript. Finally, we would like also to express gratitude to two anonymous reviewers for their insightful inputs. GD was funded by the Dean's Postdoctoral Fellowship at Lee Kong Chian School of Medicine at the time of writing early drafts of this article.

REFERENCES

Altman, Russ B. (1998). "Editorial: A Curriculum for Bioinformatics: The Time Is Ripe." *Bioinformatics* 14, no. 7, 549–550.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10). https://doi.org/10.1186/gb-2010-11-10-r106

Bartlett, Andrew, Jamie Lewis, and Matthew L. Williams. (2016). “Generations of Interdisciplinarity in Bioinformatics.” *New Genetics and Society* 35 (2). Taylor & Francis: 186–209. doi:10.1080/14636778.2016.1184965.

Bartlett, A., Penders, B., & Lewis, J. (2017). Bioinformatics: Indispensable, yet hidden in plain sight? In BMC Bioinformatics (Vol. 18, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12859-017-1730-9

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10). https://doi.org/10.1088/1742-5468/2008/10/P10008

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). Routledge.

Bogen, James, and James Woodward. 1988. “Saving the Phenomena.” *The Philosophical Review* XCVII (3).

Calder, N., Akdag, M., Aaron Press, D., & Davies, A. (2021). The Future of Clinical Bioinformaticians in the NHS: An Assessment Report and Recommendations to Build and Boost the Future Workforce.

Chari, T., Banerjee, J., & Pachter, L. (2021). *The Specious Art of Single-Cell Genomics*. https://doi.org/10.1101/2021.08.25.457696

Craver, Carl, and Lindley Darden. 2013. *In Search of Mechanisms*. Chicago: The University of Chicago Press.

Davison, Daniel B. et al. (1994) “Whither Computational Biology.” *Journal of Computational Biology* 1, no. 1, 1–2.

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*. https://doi.org/10.1186/1471-2105-7-3

Fitch, W. M., & Margoliash, E. (1967). Construction of Phylogenetic Trees. In *Nat. Phys. Lab. G. Brit. Proc. Symp. No* (Vol. 17). Academic Press. https://www.science.org

Frigg, R., & Hartmann, S. (2020). Models in Science. In *Stanford Encyclopedia of Philosophy*

Gaidatzis, D., Burger, L., Florescu, M., & Stadler, M. B. (2015). Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nature Biotechnology*, *33*(7), 722–729. https://doi.org/10.1038/nbt.3269

Gentleman, Robert C., Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, et al. (2004). "Bioconductor: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology* 5 (10).

Germain, P. L., Vitriolo, A., Adamo, A., Laise, P., Das, V., & Testa, G. (2016). RNAontheBENCH: Computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Research*, *44*(11), 5054–5067. https://doi.org/10.1093/nar/gkw448

Gorin, G., Fang, M., Chari, T., & Pachter, L. (2022). RNA velocity unraveled. *PLoS Computational Biology*, *18*(9). https://doi.org/10.1371/journal.pcbi.1010492

Gorin, G., Vastola, J. J., Fang, M., & Pachter, L. (2022). Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-34857-7

Gorin, G., Vastola, J. J., & Pachter, L. (2023). Studying stochastic systems biology of the cell with single-cell genomics data. *Cell Systems*, *14*(10), 822-843.e22. https://doi.org/10.1016/j.cels.2023.08.004

Grabowski, P., & Rappsilber, J. (2019). A Primer on Data Analytics in Functional Genomics: How to Move from Data to Insight? In Trends in Biochemical Sciences (Vol. 44, Issue 1, pp. 21–32). Elsevier Ltd. https://doi.org/10.1016/j.tibs.2018.10.010

Gray, J. M., Harmin, D. A., Boswell, S. A., Cloonan, N., Mullen, T. E., Ling, J. J., Miller, N., Kuersten, S., Ma, Y. C., McCarroll, S. A., Grimmond, S. M., & Springer, M. (2014). SnapShot-Seq: A method for extracting genome-wide, in Vivo mRNA dynamics from a single total RNA sample. *PLoS ONE*, *9*(2). https://doi.org/10.1371/journal.pone.0089673

Guala, F. (2002). Models, Simulations, and Experiments. In L. Magnani & N. Nersessian (Eds.), *Model Based Reasoning: Science, Technology, Values* (pp. 59–74). Kluver Academic.

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., & Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, *13*(10), 845–848. https://doi.org/10.1038/nmeth.3971

Huber, Wolfgang, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, et al. (2015). "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2). Nature Publishing Group: 115–21. doi:10.1038/nmeth.3252.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, *5*(9). https://doi.org/10.1371/journal.pone.0012776

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., & Morris, S. A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, *614*(7949), 742–751. https://doi.org/10.1038/s41586-022-05688-9

Keller, Evelyn Fox. 1983. *A Feeling for the Organism - The Life and Work of Barbara McClintock*. W.H. Freeman and Company.

Knorr-Cetina, Karin. 1999. *Epistemic Cultures*. Cambridge, MA: Harvard University Press.

La Manno, G., Soldatov, R., Zeisel, A. et al. RNA velocity of single cells. *Nature* 560, 494–498 (2018). https://doi.org/10.1038/s41586-018-0414-6

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. In *Genome Biology* (Vol. 15). http://genomebiology.com/2014/15/2/R29

Leonelli, Sabina. 2016. *Data-Centric Biology*. Chicago: University of Chicago Press.

Lewis, Jamie, and Andrew Bartlett. (2013). "Inscribing a Discipline: Tensions in the Field of Bioinformatics." *New Genetics and Society* 32 (3): 243–63. doi:10.1080/14636778.2013.773172.

Lewis, Jamie, Andrew Bartlett, and Paul Atkinson. (2016). "Hidden in the Middle: Culture, Value and Reward in Bioinformatics." *Minerva* 54 (4). Springer Netherlands: 471–90. doi:10.1007/s11024-016-9304-y.

Li, Y., Ge, X., Peng, F., Li, W., & Li, J. J. (2022). Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biology*, *23*(1). https://doi.org/10.1186/s13059-022-02648-4

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0550-8

Lowrie, Ian. (2017). "Algorithmic Rationality: Epistemology and Efficiency in the Data Sciences." *Big Data and Society* 4 (1). SAGE Publications Ltd. doi:10.1177/2053951717700925.

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., & Theis, F. J. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, *19*(1), 41–50. https://doi.org/10.1038/s41592-021-01336-8

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, *18*(9), 1509–1517. https://doi.org/10.1101/gr.079558.108

Markowetz, F. (2017). All biology is computational biology. PLoS Biology, 15(3). https://doi.org/10.1371/journal.pbio.2002050

McIlroy, M. D., Pinson, E. N., Tague, B. A. "Unix Time-Sharing System Forward". (1978). *The Bell System Technical Journal. Bell Laboratories*. 57 (6, part2). p.1902.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861

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

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

Morgan, M. (2003). Experiments without material intervention - model experiments, virtual experiments, and virtually experiments. In H. Radder (Ed.), The Philosophy of Scientific Experimentation (pp. 216–233). University of Pittsburgh Press. Guala, F. (2002). Models, Simulations, and Experiments. In L. Magnani & N. Nersessian (Eds.), Model Based Reasoning: Science, Technology, Values (pp. 59–74). Kluver Academic.

Mueller-Wille, Staffan, and Hans-Jorg Rheinberger. 2012. *A Cultural History of Heredity*. University of Chicago Press.

Nitzan, M., Karaiskos, N., Friedman, N., & Rajewsky, N. (2019). Gene expression cartography. *Nature*, *576*(7785), 132–137. https://doi.org/10.1038/s41586-019-1773-3

Parker, W. S. (2008). Franklin, Holmes, and the epistemology of computer simulation. International Studies in the Philosophy of Science, 22(2), 165–183. https://doi.org/10.1080/02698590802496722

Parker, W. S. (2009). Does matter really matter? Computer simulations, experiments, and materiality. Synthese, 169(3), 483–496. https://doi.org/10.1007/s11229-008-9434-3

Peschard, I., & van Fraassen, B. (Eds.). (2018). *The Experimental Side of Modeling*. University of Minnesota Press.

Ratti, E. (2020). What kind of novelties can machine learning possibly generate? The case of genomics. *Studies in History and Philosophy of Science Part A*, *83*, 86–96. https://doi.org/10.1016/j.shpsa.2020.04.001

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., … Yosef, N. (2017). The Human Cell Atlas. *ELife*, *6*. https://doi.org/10.7554/eLife.27041

Rheinberger, Hans-Jorg. 1997. *Toward a History of Epistemic Things: Synthetizing Proteins in the Test Tube*. Stanford University Press.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., & Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. In *BMC Genomics* (Vol. 13). http://www.biomedcentral.com/1471-2164/13/484

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, *14*. https://doi.org/10.1186/1471-2105-14-91

Soneson, C., Srivastava, A., Patro, R., & Stadler, M. B. (2021). Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLoS Computational Biology*, *17*(1). https://doi.org/10.1371/journal.pcbi.1008585

Stephenson, E., Reynolds, G., Botting, R. A., Calero-Nieto, F. J., Morgan, M. D., Tuong, Z. K., Bach, K., Sungnak, W., Worlock, K. B., Yoshida, M., Kumasaka, N., Kania, K., Engelbert, J., Olabi, B., Spegarova, J. S., Wilson, N. K., Mende, N., Jardine, L., Gardner, L. C. S., … Haniffa, M. (2021). Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine*, *27*(5), 904–916. https://doi.org/10.1038/s41591-021-01329-2

Stevens, Hallam. 2013. *Life out of Sequence - A Data-Driven History of Bioinformatics*. Chicago: Chicago University Press.

Strasser, Bruno. 2017. *Collecting Experiments - Making Big Data Biology*. The University of Chicago Press.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*. www.pnas.orgcgidoi10.1073pnas.0506580102

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-41695-z

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), 381–386. https://doi.org/10.1038/nbt.2859

van der Maaten, L. and Hinton, G. (2008). "Visualizing Data using t-SNE". Journal of Machine Learning Research, 9 (86) 2579-2605

Way, G. P., Greene, C. S., Carninci, P., Carvalho, B. S., de Hoon, M., Finley, S., Gosline, S. J. C., le Cao, K. A., Lee, J. S. H., Marchionni, L., Robine, N., Sindi, S. S., Theis, F. J., Yang, J. Y. H., Carpenter, A. E., & Fertig, E. J. (2021). A field guide to cultivating computational biology. In PLoS Biology (Vol. 19, Issue 10). Public Library of Science. https://doi.org/10.1371/journal.pbio.3001419

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., & Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, *20*(1), 1–9. https://doi.org/10.1186/s13059-019-1663-x

Wu, Y., Tamayo, P., & Zhang, K. (2018). Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding. *Cell Systems*, *7*(6), 656-666.e4. https://doi.org/10.1016/j.cels.2018.10.015

Xu, C., & Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, *31*(12), 1974–1980. https://doi.org/10.1093/bioinformatics/btv088

Zappia, L., Phipson, B., & Oshlack, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Computational Biology*, *14*(6). https://doi.org/10.1371/journal.pcbi.1006245

Zeisel, A., Köstler, W. J., Molotski, N., Tsai, J. M., Krauthgamer, R., Jacob-Hirsch, J., Rechavi, G., Soen, Y., Jung, S., Yarden, Y., & Domany, E. (2011). Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. In *Molecular Systems Biology* (Vol. 7). https://doi.org/10.1038/msb.2011.62