The Cancer Genome Atlas Project: Data-driven, Hypothesis-driven or Something In-Between?

Plutynski

Draft for Perspectives on the Human Genome Project and Genomics, Minnesota Studies in the Philosophy of Science
Edited by Chris Donohue and Alan Love

1. Introduction

What is the aim and character of big data sciences like cancer genomics? Some (Hey, et. al., 2009) have claimed that data-driven science is an entirely new paradigm for science – that data driven science will replace hypothesis-driven and experimental sciences. Others have proposed that there is not a dichotomy between data driven and hypothesis-driven research, but rather the approaches are 'hybridized' (Smalheiser 2002; Kell and Oliver 2003; Strasser 2012; Keating and Cambrosio 2012; Leonelli 2012).

This raises the question, however, what exactly does it mean to "hybridize" the two. As Strasser (2012, 2019) argued, when we put data driven science in historical perspective, it seems clear that the enterprise of organizing, comparing and analyzing large amounts of data has a long history, and has always required that researchers endorse hypotheses, however tentatively. From Linnaeus (Charmantier and Müller-Wille 2012) to model organism research (Leonelli, 2016), some tentative hypotheses or commitment to ontological categories is required to characterize and organize data.[1]

Here I consider the Cancer Genome Atlas (TCGA) as a case in point. I argue that this episode in the history of genomics provides further grounds to contest the sharp contrast between purportedly "hypothesis free" or "data led" and hypothesis driven research. The enterprise of collection and cataloguing genomic data in TCGA involved refinement of methods for identification and classification, a hypothesis-driven and experimental process.

[1] Ratti (2015) distinguishes two kinds of research in cancer genomics: merely descriptive data mining, versus "mechanistic" hypothesis testing. The latter of the two is explanatory, in virtue of "identification of a mechanism." Contra Ratti, in my view, at least in TCGA, the iterated and partial ways in which mechanisms are identified over time, suggests that any divide between the "merely descriptive" and mechanism identification is blurred.

Moreover, I will argue that there are distinctive features of this case, which repay closer scrutiny. TCGA was conducted in a context where there was significant pressure to produce publications, and demonstrate potential clinical applications. Grant proposals and public statements about TCGA framed the project as in service of cancer medicine: developing better prognostic and diagnostic tools, enabling "targeted" therapies. There was thus a premium on identification of "actionable" variants, efficient completion of the project, and a competitive aspect of the research. Yet, what count as "actionable" variants, and how we demarcate "driver" genes from "passenger" genes, was an open question at this stage, and one that TCGA itself aimed in part to resolve.

The project was jointly run by both NCI (National Cancer Institute) and NHGRI (National Human Genome Research Institute), which are two institutional cultures – one more "applied" and one more "basic" science. The project managers anticipated challenges in collaboration, calibration of tools, and analysis of results, in part because they were some of the same project managers of the human genome project (HGP), and so planned in advance for exactly such challenges. In light of these anticipated challenges, they modeled TCGA on the human genome project (HGP). Like the HGP, continued support and participation in the program was contingent on efficiency, quality and consistency. Project managers put the participants into competition, in order to promote productive collaboration, test out and develop novel technologies, and refine tools of analysis (Peterson, J. 2018, NHGRI workshop).

Like the HGP, TCGA was carried out by researchers with different disciplinary backgrounds and training, as well as quite different goals, which led to occasional disagreements about priorities, data quality, and analysis. Clinicians at the front end of the research (collecting samples from patients), and data analysts at the tail end of research had to learn to communicate effectively and resolve such conflicts. This was complicated by the fact that there was a bootstrapping element to this research. Many of the questions researchers were investigating could not be clearly defined until preliminary data were in. Yet, organizing and analyzing the data effectively itself required answers to these same questions. Researchers could thus not simply "let the data lead"; analysis of data itself was a process of reframing, and refining, hypotheses about how to separate signal from noise.

In sum, TCGA was neither strictly speaking hypothesis driven, nor simply natural history; it was scaffolding future research. In this way, the project is not necessarily all that different from natural history, especially natural history with the infrastructure of museums (cf. Strasser, 2019). In the early stages of the sequencing of cancer genomes, researchers were still mapping out the sample space. This was – itself – a hypothesis driven enterprise. A major site of dispute was how best to define and identify "drivers" of cancer.[2] Indeed, this issue is still a matter of dispute, as it is now becoming increasingly clear that the results of TCGA (particularly in the first five years of its tenure) will need to be revisited. The quality of the samples and methods of both sequencing and sequence analysis, some argue, suggest that the false positive rate for the original papers may have been as high as 40% (Shi, et. al., 2018).

Of course, this will not sound all that surprising to those familiar with the fact that science is a process requiring lots of trial and error. Before we can know how to test hypotheses, we need to calibrate our instruments, we need to sample the space, and we need to understand the particular challenges that face classification in some domain. As I will document below, researchers intended the process to be the first in an iterative series of steps at better mapping out cancer's genomic diversity, or, to help us identify both the "known knowns, known unknowns and unknown unknowns" (Govindan, 2018).

TCGA's major successes were the generation of tools for analysis, institutional relationships, and methodological strategies necessary for future development of truly (clinically) useful cancer genomics. As Biden might have put it, TCGA was setting the stage for the "moon shot" of precision medicine. It took us part of the way there, but it also provided important clues for exactly how far we need to go. TCGA's "marker" papers containing "consensus genomes" of the acute myeloid leukemia (or, AML), breast, prostate, and lung adenocarcinoma, published in *Cell* and *Nature* were substantive scientific achievements, but they were also tentative and preliminary. Mapping the unknown and scaffolding future research are, however, both important parts of successful science. In this way, the TCGA was not altogether different from Alan Boyden's blood

[2] The concept of a cancer "driver" is contested, but at least at first, the idea was that there were 5-8 mutations per cancer, associated with capacities typically associated with growth and successful survival of cancer cells: evading immune detection, attracting a blood supply, replicating without limit, invading neighboring tissue, resisting apoptosis, and so on. See below for further discussion.

banks, or Margaret Dayhoff's *Atlas of Protein Sequence and Structure* (cf. Strasser, 2019).

2. Background: The Planning Stage

The predecessors of this project began almost two decades ago. Li Ding recalls the initial idea developing around 2003, and the first iteration of it was called the "tumor sequencing project":

> Ding: So, I think around 2003, 2004, this idea of doing cancer sequencing came up… the tumor sequencing project. It was launched by the National Human Genome Research Institute, at that time lead by Francis Collins. This project involved the sequencing of 188 patients with lung adenocarcinoma. We decided to use a targeted gene approach. So, by working with Dr. Michael Meyerson from the Broad Institute, we came up with a 623 gene list for this project. From today's view, it is small. But, at that time, it was definitely the biggest endeavor in cancer genomics.
>
> AP: So you weren't sequencing whole exomes, or whole genomes, at the time? You were just sequencing these targeted genes?
>
> LD: Yes. 623 genes for 188 patients with lung adenocarcinoma. To get that project going, it required collaboration across the three major sequencing centers. So, that was Wash U, Broad, and the Baylor College of Medicine… So, there are a few challenges to do such a big project: Number one: capacity. When you think about 188 patients, across 623 genes. You are thinking 188 patients x2, because you have to sequence tumor and normal. And you are talking about 623 x10 amplicons, because for each gene, there are multiple exons, and we have to design PCR primers to amplify these regions, because at that time we were still using Sanger sequencing. So, we have to … there are a lot of steps involved. We need to primer picking pipeline, to ensure that we get success. And, we don't always get what we want on the first try. So we have to go back and redo, and so on and so forth, to get good coverage across 623 candidate cancer genes.

> … Eventually we were able to publish this paper in *Nature*, in 2008. It was almost a 4-5 year project, from the design to the completion. And, with I believe, if I remember right, people from 19 different institutions, under the leadership of NHGRI – at that time, Francis. NCI director, Harold Varmus was also a coauthor on this paper. He was quite interested in this project – in lung cancer. He's still interested in lung cancer.
>
> AP: At that time, were you thinking about a longer term, or were Collins and Varmus already thinking of a longer-term project – of looking at whole genomes? Or…?
>
> LD: This project *was the pilot* for TCGA. This project was sponsored by NHGRI. So, with the initiation of this project, the leadership at that time has already started to think about doing more. That's why TCGA was launched… this project – the tumor sequencing project – really at first was a *technical exercise to understand if we could do this at a larger scale* …the very first TCGA project on glioblastoma *adopted the exact same strategy, or project design*. So, we did 200 tumors, across about 600 genes, using targeted sequencing, using Sanger approach. So, that paper was also published in 2008. Because we learned so much from the TSP, we were able to push the GBM [glioblastoma] a little faster. (Ding, italics added, October 2018)

In other words, TCGA was itself the second generation, the second stage in a long-term plan on NHGRI's part to invest in large scale cancer genomics. And, TCGA itself was an iterated process, one that became progressively more sophisticated, incorporating newer, faster technology, and requiring the coordinated efforts of several institutions.

The Cancer Genome Atlas program (TCGA) was a federally funded U.S. research program, supported jointly by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). Beginning with a pilot grant of $100 million in 2006, the aim initially was to sequence the genomes of three major cancers: ovarian, lung and brain. With the advent of next generation sequencing, and Obama's American Recovery and Reinvestment Act, following the economic downturn of 2008, the program expanded. By 2016, 33 cancers had been sequenced, based on over 11,000 samples. The

program was officially concluded in 2016, at which point, a total of 18 analyses of individual cancer types had been published in landmark papers in *Nature* and *Cell*.

In 2007, then director of the NHGRI, Francis Collins, described the goal of the TCGA as "…identify all the genetic alterations in different forms of cancer so that gene changes driving the disease can be targeted directly… As each new type of cancer is studied and added to TCGA, investigators will gain another rich new set of genomic targets and profiles that can be used to develop more tailored therapies." (Collins & Barker, 2007) In addition to molecular classification of different tumor types, the hope was that scientists would identify novel mutations associated with different cancer types and subtypes, enabling better risk analysis, and prognostic tools, as well as better appreciation of mutation rates across cancers, and the extent and nature of heterogeneity both within and across different tumor types, as well as shared pathways affected. Collins and Barker pointed to examples of successful targeted therapies, such as Gefitinib, as having the potential to grow out of this research.

Examples like Gefitinib, an EGFR inhibitor, used in breast and lung cancers, are often taken to be exemplars of the great potential benefits of "genomic" driven medicine. However, this is somewhat of a rewriting of history, since many of the drugs identified as "precision" therapies were approved for clinical applications decades prior to any efforts in cancer genomics, let alone TCGA. For instance, Herceptin (or, traztuzumab), a drug that targets receptors in Her2+ breast cancer, was approved by the FDA in 1998. Some of the major signal transduction and cell-cycle inhibitors were developed and approved by the FDA well before the TCGA. Gefitinib was approved in 2003.

Inquiry into the causal role of various mutations and pathways in cancer was underway long before cancer genomics.[3] Indeed, the TCGA is one of the latest of several internationally funded research programs in cancer genomics. The Catalogue of Somatic Mutations in Cancer (COSMIC), a database of genetic mutations associated with cancer, hosted by the Sanger Institute in the UK, has been in operation since 2004. The International Cancer Genomics Consortium has been ongoing since 2006. TCGA was a

[3] Why these drugs were effective (relatively speaking) and others not requires a much lengthier analysis than is possible here. For further discussion of the scope and limits of precision oncology, see, e.g., London, et. al., 2019; Kimmelman, et. al., 2015, 2018; Hey, et. al., 2016; Tannock, et. al., 2017; Tao, et al., 2018; Prasad, et. al., 2016; Schrager, et. al., 2019; Plutynski, forthcoming.

part of this international consortium, though it was one of the largest projects underway at the time, internationally.

While the goal most widely advertised of the TCGA was to better diagnose and treat cancer, there were in fact many different kinds of goals at play in the TCGA. Alongside the potential scientific and clinical applications, the TCGA promoted improvements in speed, accuracy, and lower cost of sequencing, leading in part to the drastic reduction in cost of sequencing an entire genome – from as much as $100 million in 2001, to as little as $1,000 in 2016 (Hutter, 2018). Promoting and developing productive interactions between the scientists and the biotech facilities developing these technologies were central aims of the project. That is, there were economic and institutional drivers motivating TCGA. It is not coincidental that the project expanded and was completed shortly after Obama's Recovery act, a central goal of which was investment in science and technology. The goal of TCGA was not simply scientific understanding. To be sure, the participants hoped for a better understanding of cancer, but also saw their project as promoting the development of biotechnology, and of new kind of interdisciplinary, team science, development of better methods of sequencing, analysis, curation, storage, accessibility and sharing of genomic data. While the latter is surely a means to the former, it was not merely in service of the former. The latter goals stood on their own as major achievements – among the first mentioned in an interview with the program's director for the last five years of its tenure (Hutter, 2018).

Ultimately, however, the ostensible aim of the project was to complete genome sequences of 33 different cancer types, taking samples of each cancer, paired with normal or healthy tissue samples, which served as a "reference" genome for comparison (Hutter, 2018). The TCGA led to several striking insights into the etiology and heterogeneity of cancer. For example, sequencing of the breast cancer genome led to novel subclassifications of breast cancers into four major molecular types (Cancer Genome Atlas Network, 2012), and comparisons across diverse cancer types led to the realization that some cancers arising in different cells of origin shared more molecular similarities than cancers arising in the same cells of origin (Hoadley, et. al., 2014). The TCGA's results were made publicly accessible via a portal developed at UC Santa Cruz, the Cancer Genome Hub. Developing accessible public cancer genome databases was a key

element of the program. And, the future was always in view; a second generation "Pan-Cancer Atlas" project has already mined this data to produce papers on cancer classification, overlapping pathways or common "oncogenic processes" across a variety of cancers, and a variety of signaling pathways associated with cancer (Hoadley, et. al., 2018; Malta, et. al., 2018; Ding, et. al., 2018).

There were many surprises that came as a result of TCGA, but one issue that was never in question, one on which the project was founded, was that cancer is "a genomic disease."[4] Collins and Barker open their article promoting TCGA with a quotation from Dulbecco, "if we wish to learn more about cancer, we must now concentrate on the cellular genome." As Collins and Barker explain, "Since the first identification in 1981 of a cancer-promoting version of a human gene, known as an oncogene, scientists have increasingly come to understand that cancer is caused primarily by mutations in specific genes" (Collins and Barker, 2007, p. 52).

The ontological commitments or theoretical presuppositions standing behind TCGA were that mutations and chromosomal changes are major causes of cancer – the "causal bottleneck" through which all processes yielding cancer flow. Mutations are, in this view, centrally responsible for cancer cell survival, persistence, invasion and metastasis. If we can identify those genetic and genomic abnormalities, we may be better able to provide targeted treatments for the disease. These "driver" genes and their downstream effects on various pathways or molecular markers might serve as biomarkers for prognosis, and potential targets of therapy. Collins and Barker are explicit: if we can intervene on these, we may then disrupt the functional disruption typical of cancer cells. This explains Collins' appeal to targeted treatments such as Gefitinib; such successes were taken to be one particularly vivid source of evidence in favor of TCGA.

However, as the TCGA was going on, it became increasingly clear how factors outside of the genome play a significant role in cancer initiation, progression, and persistence of the cancer phenotype. Though researchers were aware of the role of extragenomic factors in cancer long before TCGA, it became better appreciated how (and

[4] This quotation comes from Carolyn Hutter, who explains as follows: "it actually messes up the genome quite dramatically, you know, a whole giant chromosome arms switch around." She is referring to the chromosomal abnormalities associated with cancer cells. To be sure, this claim can be interpreted in a variety of ways (Plutynski, 2018), and might commit one to very different claims in different contexts.

how much) epigenetics, extra-genomic DNA e.g., microRNA, tissue architecture, the microbiome, immune response, and metabolic factors in the extra-cellular environment act to promote or halt the progression of disease (Jones, et. al., 1999; De Visser, et. al., 2006; Esteller, 2008; Bissell and Hines, 2011; Vander Heiden, et. al., 2017; Böttcher, 2019) Indeed, the Pan-Cancer Atlas project is examining how and why these features of cancers interact in the progression of disease (Ding, 2018).

Nonetheless, there were several major theoretical presuppositions driving the methodological revisions to cancer genome analyses: (1) there is a distinction between mere "passenger" and "driver" genes in cancer, (2) "drivers" play a distinctive causal role in the initiation and progression of disease, (3) identifying these drivers will provide ways to identify optimal targets for intervention, or treatment, as well as biomarkers for patient stratification, or risk assessment in prognosis and treatment decisions, and (4) there are many genes already known to be associated with cancer, which we ought to find representatives of in our samples of various tumor types and subtypes. Each of these assumptions were in play, both in driving what was expected to be discovered by the sequencing efforts, and in interpreting what was discovered. Attention to the tissue architecture, its influence on mutation rates typical of the diverse cancer types, or role in shaping tumor microenvironment, was just not on the table as relevant to the analysis.

While the initial goal was 500 samples per major cancer type, in some cases (for rare cancers, or cancers where storage or access was exceedingly difficult or expensive), fewer samples were identified (50-100), due to difficulty in acquisition of samples of these rare diseases, or poor quality of samples, in some cases.[5] According to Carolyn Hutter (Hutter, 2018):

> Initially, the idea was to do twenty cancers, five hundred cases per cancer. And, to then to get to 10,000. And, the truth of the matter is that sample acquisition is challenging. You know, you needed to have samples that had the right quality, the

[5] Different quality samples has to do with whether the biopsies were gathered, tracked, and stored in consistent ways, and had more or less non-cancer cells included (stroma are the tissue surrounding the cancer, and affect the analysis), among many other factors. The standards for storage and analysis of tumor biopsies themselves evolved as TCGA went on. Initially biopsies were Formalin-fixed paraffin-embedded (FFPE). However, these were eventually considered an unreliable source for gene expression analysis due to the partial RNA degradation. Fresh-frozen (FF) samples were later found to be more reliable and largely unaffected by the storage time.

> right consent, all of the details that came along with it… in some cases, getting 500 qualified cancers could happen. And, in some cases getting 500 qualified cancers didn't happen. And, so, there were some changes. One of the changes that did happen was the rare cancer project. So, upping the number… But, then when we went for rare cancers – you know, the target for rare cancers was really the 50-100 range. (Hutter, interview, August 2018)

The "characterization" (or, analysis of the genomic data gathered) was completed "on time," in 2015, and the last marker paper – a consensus statement of the genomic features of each cancer type – published in 2016. These papers were the central publications growing out of TCGA: marking completion of the breast, thyroid, prostate, etc., genome. Though, the "completeness" of the analysis was only in light of agreed upon standards, which shifted as the process evolved.

To clarify, cancer genomes are consensus objects – samples of dynamic, heterogeneous classes of disease processes. That is, each cancer is a population of cells that changes over time, acquiring novel mutations as the disease progresses (Greaves, et. al., 2010; Frank, 2007). How and how much cancers change over time was – at least initially – unknown. Initially, on the supposition that mutations of major effect both initiated cancer, and were retained throughout this dynamic process, samples were taken at first diagnosis. However, as TCGA went on, it became increasingly clear that tumors might become more heterogeneous over time, acquiring changes may play important roles in invasion and metastasis. Thus, while initially samples taken at first diagnosis seemed appropriate, the current view is that for genuinely "targeted" therapy, a patient should have samples taken at several stages in the progression of the disease (Lee, 2019; Gyanchandani, 2018).

In other words, the completed genome at publication was founded on consensus sampling standards, which were context sensitive in two ways: (1) with respect to the current understanding of the disease and methodological/technological advances; and (2) with respect to distinct goals – scientific and practical. For instance, the depth of reads (how many copies of sequence one generated to line up sequences) was 30X at the beginning of the project. Today, some researchers estimate that for cancer, the depth of

reads should be as high as 150X, if we wish to get accurate measures of the most significant genes playing a role in the disease (Griffith, et. al., 2019).[6]

In sum, the TCGA was in service of many (and open-ended) goals, and standards of achievement of these goals changed as the technology and knowledge advanced. Cancer genomics is in this way a kind of "discovery" science – in part, confirming and deepening understanding of what we already knew, but also raising many questions, suggesting novel hypotheses, and forcing refinement and reframing of goals. Stumbling blocks or challenges were to be expected. Below I consider a concrete example of one challenge that arose in the analysis of cancer genomes, but first, I offer a bit more background on the nature of the project itself.

3. Scientific Background: What Does it Mean to Sequence a Cancer Genome?

There are (roughly) two steps along the way to the generation and analysis of a consensus cancer genome. First, "mutation calling" is a technique whereby one identifies the mutations in a single cancer; or, a mutation caller is a "classifier asking at every locus, 'Is a mutation here?'" (Getz, 2015). Mutation calling involves identifying which variants are unique to a specific patient, and which are mere artifacts. The estimate depends on a variety of factors: the allelic fraction (how much of the DNA in a sample comes from tumor cells, and how much from healthy cells), the extent of "coverage" of sequence of both tumor and normal (reference) genome, and the extent of noise introduced in sequencing alignment methods. The aim is to represent the likely number of "hits," or "driver" mutations. Li Ding explains as follows:

> … we're comparing the cancer sequence in the tumor cells to the normal cells in that same individual. And you can almost line them up and see where do they differ. And, where they differ is a mutation. So, mutation calling is the act of lining them up and seeing where they differ. Now, if genomes were ten base pairs long, that would be easy. It's like, here's my ten base pair normal, here's my

[6] This is something specific about cancer that demands an increased depth of reads (30X is adequate for normal genomes). What's distinctive about cancer genomes is that they are more heterogeneous, and so it's important to have a higher replication of alignments to ensure that the reads are not false positives.

> normal, where do they differ. But, mutation calling becomes more complicated because you have to then make these calls based on the sequence that you have, and so you have to make sure: "Did I accurately call the normal sequence correctly? Did I accurately call the tumor sequence correctly?" If it's a base pair mutation, sometimes that's easy to do. But, sometimes we have…
> [AP: Inversions? Deletions?]
> Ding: Right. And, what happens in cancer … cancer is a disease of the genome, and it actually messes up the genome quite dramatically, you know, a whole giant chromosome arms switch around. So, you have to reconstruct these large scale and small scale structural variations. You have to reconstruct all of that to even be able to line it back up – to compare it… So, mutation calling in the somatic contexts becomes taking the information you have about the germline, taking the information you have about the tumor sequence, and identifying what are the differences. What has happened? How does the tumor sequence look different from the germline sequence? So, there's different programs, and different mutation callers that people have used to do that, for structural nucleotide variants, for structural variations, for larger rearrangements. And, they're just computationally slightly different and they have different assumptions, etc. And, what's generally been found has been that the best thing to do is use more than one approach and come to a consensus call. (Ding, 2018)

Typically several different mutation calling methods are used, and a consensus is arrived at. That is taken to be the "consensus" found in a particular individual. This data from each paired normal-tumor sample is then compiled and analyzed to generate the "molecular landscape" of a cancer. These are the "driver" mutations taken to be typically associated with all cancers of a specific type or subtype: prostate, breast, etc.. Algorithms are used to estimate their significance, in part in light of relative frequency, but in part in light of data already available from cell and molecular biology about the causal roles of specific genes in cancer, what are called "candidate" drivers.[7] This process of using

[7] Candidate drivers are either well-known to be associated with the cancer phenotype, e.g., they play important roles in the activation of the cell cycle (birth and death) (e.g., TP53), cell division and repair

algorithms to identify driver genes is called the 'characterization' of a cancer genome – yielding the "consensus" genome published in the marker papers. Each cancer genome was analyzed using a variety of evolving, ever more refined tools for aggregating the data generated from individual sequencing and mutation calling.

Prior to second generation sequencing, this was done with a baseline assumption of average mutation rate per cancer type, which was itself averaged based on the sample taken (cf. Lawrence, 2010, p. 2). This assumption, however, led to some difficulties. What needed to be considered – in addition – was the variation of mutation rates within and across cancers, as well as the heterogeneity in type of mutation and mutation frequency in different parts of the genome. It was only after the cancers had been sequenced that analysts could know, as a matter of fact, how heterogeneous the mutational landscape of cancer is.

Indeed, arguably, those who initially planned and designed TCGA couldn't have known exactly how many samples to take of each cancer without appreciation of the mutation rate and heterogeneity of mutation dynamics across cancers. It was something that needed to be learned – in a boot strapping fashion – by carrying out the process of sequencing, itself. In response to persistent queries about how they determined appropriate sample size for the AML genome, Ley explains:

> No one knew what the sample size needed to be when we started. No one had a clue. We did not know the mutation rates for AML samples going in. So, the estimate of sample size was done post hoc… We looked at the first couple of dozen cases had been done… and our idea was simple: can we model the number of cases that we'd need to sequence to identify 95% of the mutations that occur in

(BRCA I, II), based on prior work in cell and molecular biology, as well as family history (e.g., APC was discovered in part, in light of mutations in this gene associated with Li Fraumeni syndrome), or, any gene that appears mutated at high frequency in many cancers, though high frequency per se is not viewed as sufficient to count a gene "in." Second and third generation algorithms were developed that took into account not only frequency, but also "built in" information about relevant functional role – genes were weighted more heavily that were known to play roles in cell birth, death, or accuracy of replication. As the process went on, researchers identified several mutations that were arguably false positives. For instance, being a large gene and thus a large target, or, more subject to mutations due to the timing of its replication during the cell cycle, etc. increasingly were ruled out. See below for further discussion.

> at least 5% of patients? Based on the sequencing of the first 24 AML genomes, we calculated that number to be 200, and it was on the money. (Ley, July 2018)

In response to similar questions about sample size for the lung cancer genome characterization, Govindan (who worked on the lung cancer team) explains that pragmatic factors played a role:

> … the sample sizes are somewhat arbitrary… In fact, I can speak with more confidence about lung cancer. The main thing is… for lung [cancer], there was a decision made to do 500 / 500 [each of each type of lung cancer]. It's a round number. There are two common types of lung cancer: adenocarcinoma and squamous cell. So, you know, 500/500. … 500 seemed like a doable number. So, that's how the decisions were made to the best of our knowledge. I led the TCGA lung cancer project… but things were beyond our decision making, and the NCI did that. Also, remember, the cost of procuring specimens – not just the cost of sequencing – was pretty high at the time… The money factors in too. So, they had a bunch of money, and they wanted to sequence some tumors, and they had to make a decision. (Govindan, July, 2018 (inserted parentheticals are mine))

In fact, one of the earliest challenges in TCGA was obtaining high quality samples. Initially, samples of glioblastomas were not of sufficiently high quality to generate adequate data for sequencing (Peterson, pers. com., 2018) It was this matter of quality and annotation that was most on the minds of researchers at the beginning stages, not necessarily quantity of samples. This, in addition to adequate annotation of patients, turned out to be one of the most time consuming (and not coincidentally, expensive) aspects of the program. Tim Ley commented, as follows:

> One of the great myths of cancer genomics is that grants usually pay for the collection and clinical annotation of samples, in addition to sequencing and interpretation. The sequencing is now less expensive than annotating the samples and the information about the patients.. It costs us far more money to bank and

> annotate AML samples than it does to sequence their exomes. … At any given time, we are following hundreds of patients with acute leukemia or myelodysplastic syndromes so that understand outcomes for every patient. The costly part now– the part that tends to be underfunded– is collecting the samples and annotating the cases…(Ley, May 30, 2018)

So, at the outset, TCGA participants were primarily concerned with quality of samples. Concerns about sample size were less central, in part because it was not even clear at that point what an adequate sample size might be, and in part because the assumption was that more and better sequencing would be possible as expense and technology improved, leading to ever larger and more comprehensive samples of similar cancer types. Indeed, Ding argues that there were several phases of TCGA, with early phases more like an extended pilot, experimental phase, and the latter stages built on this:

> I define TCGA as three phases: the early phase was GBM and ovarian, so from 2005 to 2010. Then between 2010 and 2016, I call that the production phase. TCGA was in a very stable productive phase, and we were able to publish multiple marker papers per year.
> … when you think about it, the early phase of TCGA: we spent… For example, for the GBM project, the entire consortium worked on one project, one tumor type, one paper, for almost four years. But we were *busy*, because we were trying to figure out how to do this right. Project design: how to get patients' samples in place. We had a lot of conference calls, a lot of meetings to debate about how to do this right. Sometimes it could be heated discussion, but very productive, because we need to hear different opinions … how to do this project right. Without these discussions, we weren't going to have this nice smooth signaling production phase. Then, for ovarian cancer like I said: we threw all the technology at this cancer type. For this sequencing project was a mixed bag. But I think it was very helpful for the consortium, because if we did not do ovarian cancer, we wouldn't know which was the best strategy for moving forward for the rest. So, I think those two projects were very important for TCGA. That is why I

> call that early phase TCGA. Then, we got into this nice stable productive phase, generating multiple papers a year. (Ding, 2018)

The goal was, at least insofar as TCGA was concerned, to arrive at a preliminary "consensus" sequence by compiling mutation calling data on many different patients – how many were necessary was at this point an open-ended matter. How then was the data compiled to generate a consensus sequence? When and why were consensus sequences agreed upon?

The answer depends upon the date of publication of the consensus sequence. That is, the grounds for consensus – depending as it did on evolving criteria of adequacy of samples, as well as methods of analyses and the details of the cancer type or subtype – were not fixed. Methods of reaching consensus were refined as the process went on – from the first consensus sequence (of glioblastoma) to the last (adrenocorticalcarcinoma). Different genomes of different cancer types relied upon new methods of analysis as they became available, often arriving at a consensus given overlapping results from several different methods.

4. Stumbling Blocks: the Karenina Paradox

Although there are many challenges facing analysis of cancer genomes, the focus here will be on one particular challenge, arising from the distinctive heterogeneity of cancer genomes. In a 2010 commentary in *Nature*, Meyerson, et. al., nicely sums up this challenge in his paraphrase of Tolstoy's Anna Karenina:

> … normal human genomes are all alike, but every cancer genome is abnormal in its own way. Specifically, cancer genomes vary considerably in their mutation frequency,… in global copy number or ploidy, and in genome structure. These variations have several implications for cancer genome analysis: the presence of a somatic mutation is not enough to establish statistical significance as it must be evaluated in terms of the sample specific background mutation rate, which can vary at different types of nucleotides ... The analysis of mutations must also be

> adjusted for the ploidy and the purity of each sample and the copy number at each region. (Meyerson, et. al. 2010)

To clarify, at the time Meyerson and colleagues' paper came out, the first round of sequencing data had just come in, and the first "pilot" genome, glioblastoma, had been sequenced, analyzed, and the consensus genome published. Data was already accumulating on AML, ovarian, and lung cancers. Meyerson and colleagues were reflecting on the novel task of analyzing what Ley called a "firehose" of data (Ley, 2018). There was a great deal more complexity and diversity to the information than anyone had anticipated, and the challenge was to separate signal from noise, or to identify the most significant mutations associated with cancer types and subtypes.

Prior to TCGA, the assumption was that there would be about five to eight major mutations per cancer type that played a significant role in capacities typically associated with growth and successful survival of cancer cells: evading immune detection, attracting a blood supply, replicating without limit, invading neighboring tissue, resisting apoptosis, and so on (Ley, 2018). These were the "driver" genes. While the origins of this term are somewhat obscure, Stratton et. al. (2009) use the term "driver" to refer to mutations that "confer growth advantage on the cells carrying them and have been positively selected during the evolution of the cancer." In contrast, mere "passengers… do not confer growth advantage, but happened to be present in an ancestor of the cancer cell when it acquired one of its drivers" (2009, 722). That is, Stratton seemed to assume that functional role and the selected effects of drivers would necessarily coincide, but they need not. Estimates of the number of drivers typical of each cancer type were drawn from patterns of age of incidence and were believed to range from 5-7 mutations per adult epithelial cancers, though it was believed that haematological cancers may require fewer. That is, waiting time to cancer was clearly a function of age, and so it was believed that cancer was a product of the cumulation of rate-limited mutation events (Frank, 2007; Miller, 1980).

In addition, experimental (knockout) studies also appeared to show that engineering changes in the functions of at least five or six genes in normal primary human cells is necessary to convert them into cancer cells (Schinzel, et. al., 2008). It was

these mutations that targeted treatment aimed to intervene upon. Such genes were "actionable," and so knowing which ones were the major players in cancer was exceedingly important. But, the very concept of a "driver" gene was in flux during this decade. Indeed, in part as a product of the process of TCGA, it became apparent that there were many more genes involved in cancer, but it was (at first) impossible to know which genes were most important. At first, the assumption was that with larger sample sizes, one would better be able to refine estimates and identify specific mutations significant for each cancer. However, paradoxically, the opposite turned out to be the case.

In part, this was because (as we have discussed), participants in TCGA were involved in ongoing refinement of algorithms and techniques used to identify drivers. Different algorithms deployed different presuppositions about the relative significance of various drivers. Moreover, technical artifacts, or results based on mistaken assumptions, tools, or techniques, were sometimes only uncovered after repeated instances of counterintuitive results. By way of example, initially some genes were identified as "drivers" that were in fact, simply larger "targets" – genes for which it was easier or more likely to acquire "hits." That is, especially large genes (with no apparent functional role in cancer) appeared more likely to acquire mutations. Govindan (2018) describes an example:

> … in the early days of cancer genome sequencing, one of the most common one that popped up was a gene called Titan. And, Titan was properly named Titan, because it was the largest gene. You know… it was mutated in every cancer. Later on, it became obvious that it was mutated because it was a large gene. So, the larger the gene is, there are more chances of it having a [mutation]… it's like you walking around with a target, and the larger you are, the more chances you will be hit. You know, it took a while to appreciate that. There are a lot of passenger mutations in cell populations that are present in the genome, and the larger the gene is, the greater the chances are going to be … [that it will be mutated]. So, then we learned to correct for gene size, and the variations of different things that came in, too… and, that took away a lot of noise. Basically,

> the first level is technical artifact … there were a lot of alterations in a gene that don't do anything… in the early part the fact of the size of the gene matters was one of the insights that came about… (Govindan, 2018)

While such difficulties as gene size serving as a confounding factor were eventually identified and corrected for, a rather different problem was the dramatic number of apparently significant mutations with increasing sample size. The expectation was that with more samples and better sequencing, a stronger signal would be present, indicating the most significant mutations in each cancer. But, with more samples, paradoxically the number of mutations associated with each cancer appeared to increase. Tim Ley describes this growing realization in the context of AML:

> … in the second genome that we sequenced, that was that whole theory sort of came off the rails for us. Because, the second patient that we sequenced was normal karyotype AML, and now we had done a whole genome with much greater detail, using paired-end reads. The sequencing technology was better, and we were a lot more confident about the coverage of the number of mutations we were detecting. So, here we were in this particular genome, and we could start ordering the events of the cancer. We could tell how many mutations were present in all the cells of the tumor. And that number was more than 500. And then we had to say to ourselves, " It's impossible that there are 500 mutations that are relevant for this cancer. So, was kind of a moment of truth for us, and how people think about how passenger mutations evolve, and where these mutations come from…(Ley, 2018)

Ley and colleagues had assumed that very few mutations were necessary for AML; hematopoetic cancers by and large were assumed to be associated with far fewer mutations. Thus, most of these 500 mutations had to be "noise" or "passengers" – alterations in the genome of cancer cells that play no role whatsoever in cancer's

etiology.[8] Ley (2018) describes the going alternative theories to explain the excess mutations discovered in the case of AML:

> …there are only a limited number of things this could be:
> - One possibility is that this is a "big bang" in one cell… there's an explosion of mutations that occurs in one cell at one moment in time. They all move forward together as an entity of one, and something about that cell acquiring all those mutations on that one day… that "experiment of nature" succeeds, because there are several mutations that are relevant that occur to that cell. Maybe this is a process that is going on all of the time—but you only detect the events that are successful in producing a new cell that has a significant growth advantage. And we call this AML.
> - Another possibility that was unsatisfying was the possibility that it really was an evolutionary process, but it took 500 events. Nobody liked that idea. That just seemed, like… How would you ever get a cancer if there were five hundred events that had to occur for it to be successful?
> …The third possibility (which is probably the correct one) was that nearly all of the detected mutations had to antecede the relevant mutation. They had to occur over time. And, they were random mutations that were occurring as a function of time. They were falling in places that were irrelevant for the function of the transformed cell, and they were essentially being "captured" by selection for the cell with the relevant mutation. (Ley, 2018)

In AML, Ley argued, these excess mutations were accumulated simply as a byproduct of errors that occurred during hematopoetic stem cell divisions, over the course of a lifetime. In a paper published in *Cell* in 2012, Ley and his colleagues John Welch and Daniel Link demonstrated that one could use the number of stochastically accumulated mutations in hematopoetic cells as a kind of molecular "clock," representing age of the

[8] To be clear, I'm not here endorsing Ley's reasoning, simply recapitulating it. Ley is assuming that all passenger mutations are in fact "noise" – that is, they are just those mutations that have no mechanistic link to the behavior of cancer cells. A mutation that does have such effects (e.g., a mutation that leads to problems with cell repair during replication, e.g., BRCA I or II) is by definition a "driver" mutation.

patient (Ley, et. al., 2012). That is, each individual has a certain number of mutations acquired over the course of their lifetime in hematopoetic stem cells.[9] These serve as markers of age, since rate of mutation per stem cell is a constant, and hemopoetic stem cells divide at a regular rate, as we age.

The question then became: did this explanation of the accumulation of mutations also explain the apparent excess in mutations found in other cancers? Could one extend this explanation for the apparently excess mutations in AML to all cancers? Were all these excess mutations simply byproducts of stem cell division across all tissues? Or, were they acquired over the course of cancer development, perhaps due to the characteristic chromosomal or genomic instability of cancer cells in solid tumors? Last but not least, which of these apparently significant mutations were either artifacts of either sampling methods, or insufficiently fine-tuned analysis of the data? In sum, this new influx of data led to a whole series of debates in the literature about the dynamics of cell division in development, and what role it played in the heterogeneity of cancer. This led to a related debate about the relative proportion of cancer risk was down to "luck" – for, cancer risk is orders of magnitude higher in some tissues and organs (e.g., epithelial cancers) than others (e.g., bone and brain). The former were, arguably, largely byproducts of mutations acquired in somatic stem cell division (for recent discussion, see, e.g., Tomsetti and Vogelstein, 2015, 2017; see also, Wu, et. al., 2016; Nowak and Waclaw, 2017).

While the relative role of chance mutations acquired during stem cell division in cancer is still contested, it was clear to almost everyone that a significant proportion of mutations identified as "cancer genes" using then-current methods were simply "passengers" – understood as mutations that appeared to play no functional role in cancer. Lawrence, et. al., write, (2012) "when we applied current analytical methods to whole-exome sequence data from 178 tumor-normal pairs of lung squamous cell carcinoma, a total of 450 genes were found to be mutated at a significant frequency... While the list contains some genes known to be associated with cancer, many of the genes seem highly suspicious based on their biological function or genomic properties."

[9] To be sure, the idea that cells acquire mutations as we age, due to stem cell division, was at that point well known. The insight of Welch and Ley was the particular link between age and mutational profile of AML.

For instance, mutations in genes associated with olfactory receptors, muscle proteins, and the Parkinson protein were not likely to be significantly associated with cancer.

Which mutations were the genuinely significant ones, and why do more significant genes appear to crop up as sample size increases? Lawrence, et. al., summed up the problem:

> … we describe a fundamental problem with cancer genome studies: as the sample size increases, the list of putatively significant genes produced by current analytical methods burgeons into the hundreds. The list includes many implausible genes (such as those encoding olfactory receptors and the muscle protein titin), suggesting extensive false positive findings that overshadow true driver events. Here, we show that this problem stems largely from mutational heterogeneity and provide a novel analytical methodology…
>
> The expectation has been that larger sample sizes will increase the power both to detect true cancer driver genes (sensitivity) and to distinguish them from the background of random mutations (specificity). Alarmingly, recent results appear to show the opposite phenomenon: with large sample sizes, the list of apparently significant cancer genes grew rapidly and implausibly… We hypothesized that the problem might be due to heterogeneity in the mutational processes in cancer. While it is obvious that assuming an average mutation frequency that is too low will lead to spuriously significant findings, it is less well appreciated that using the correct average rate but failing to account for heterogeneity in the mutational process can also wreak havoc… (Lawrence, et. al., 2013)

The source of the problem, according to Lawrence, et. al., was in part the extent of heterogeneity in mutation rates and mutational landscape across and within different tumor types, and in part, not taking this variation into account in the analysis of genome data. By not considering the heterogeneity of rates and types of mutation events across different cancer types in the analysis, they overshot the estimate of driver genes. Three different types of heterogeneity were relevant to the extent of false positive (and negative) results:

- Heterogeneity across cancers, and across patients of a cancer type
- Heterogeneity of mutation spectrum (rate/types of mutation, e.g., from CT, or A-G)
- Heterogeneity across the genome, in extent and character of mutations

It turned out that with more samples, the signal to noise ratio made the tests of significance too sensitive, thus identifying too many mutations as significantly associated with cancer. The extent of and nature of heterogeneity of mutation rates and total mutations across cancers complicated the analysis of the data. With mutation frequency across cancer types differing by as much as three orders of magnitude, and even within cancers by as much as two, the analysis required fine tuning tests of significance to specific cancer types and subtypes. In addition, the heterogeneity in type and location of mutation needed to be included in the analysis, in order to correct for the over-counting problem.

Without knowing what the baseline number and rate of different types of mutations across different cancers, it would be enormously difficult to determine the frequency of mutations of significance for each cancer type and subtype. But, the scientists could not have know the extent of heterogeneity in mutation rates and types across cancer without doing the sequencing in the first place. Fine tuning the methods of analysis of the data were only possible after one had the data and analyzed it (incorrectly) at first. This is a vivid example of "scaffolding" future science – or, boostrapping. Without knowing the sample space, and the character of your sampling instruments, you cannot know what and how to evaluate your results. But, you cannot know how your instruments work without taking some samples in the first place.

5. Conclusions.

The case of TCGA provides yet another instance of a challenge to the purported sharp divide between "data driven" and hypothesis driven science. Indeed, this challenge has been pointed out in other instances of big data science – from model organism research, to Dayhoff's atlas of protein sequences, to museum collections (cf. Leonelli,

2016; Strasser, 2019). TCGA was – in large part – a pilot project: they were learning how to collect, annotate, store and track high quality samples, garnering a sense of the sampling size appropriate for cancers of different types and subtypes, developing technologies and methods of analysis, building online platforms for accessing and sharing data, and coordinating the efforts of hundreds of clinicians, biostatisticians, geneticists, and computer scientists at major institutions around the U.S.. Simply learning how to gather, store, and process (more or less) high quality samples, let alone master new methods of high throughput sequencing, and analyze the results was at the core of the project for the first five years of TCGA's tenure.

The project ran into stumbling blocks, and while many were expected, some were unexpected. Learning how to coordinate research, agree upon common standards for sampling and storage of biopsies, as well as correct for batch effects in sequencing and develop common algorithms for analysis were all unsurprising. More surprising, however, was that as more samples were taken, the number of driver mutations seemed to climb, far more than was expected initially. While it may have initially seemed a paradox that there were more false positives with larger sample sizes, the paradox resolved when TCGA researchers re-thought their presuppositions about the base rate of mutations in different cell types, in different tissues. The reason for this paradoxical result, in other words, was that the researchers were making a (slightly more complex) version of the base rate fallacy. This is the fallacy of failing to take into account the "base rate," or rate at which events typically occur, in one's analysis. In estimating the number of drivers in a cancer type, they had overestimated the number of significant events because they initially underestimated the mutation frequency of particular types of genes (large genes, etc.), as well as the background rate of mutations and types of mutations in different cancer types (in cells or tissues with high rate of turnover). Incorporating this knowledge allowed them to arrive at a more conservative (and likely, more accurate) estimate of the number of driver genes typical of cancers of various types.

This example illustrates a larger, more fundamental challenge to the notion that data can simply "lead" in big data research. Genomic research such as this began with some fundamental assumptions about how and why cancer behaves as it does, assumptions that more or less reduced cancer causation to the activity of genes. But, the

assumption that a handful of major genes might be sufficient to generate future research into major pathways "for" cancer, in service of both stratifying patients and identifying therapeutic targets, became increasingly less plausible as the project wore on. It became increasingly clear that there is no direct, one-to-one causal pathway by which these genes act in the process of cancer progression. Genetic mutations are one, and only one, component, in a complex network of causal pathways, shaping the course of disease. Complex interactions with the immune system, the microbiome, tissue architecture, hormonal factors over the course of development (as well as during pregnancy and nursing), and much else, act together in modifying mutations' roles in cancer progression. Indeed, investigating these extra-genomic, interactive factors involved in gene expression and cancer dynamics has been a central goal of the next generation of the "pan-cancer atlas project": the investigation of the cancer "epigenome," "immune landscape" and so-called "cancer microbiome" (Dayson, 2017; Thorsson, et. al., 2018; Sepich-Poore, 2021). Many of the genes identified in TCGA have now been identified as either activated by the immune system, or interacting in complex ways with the microbiome, shaping disease course, or response to treatment. This realization has led to a more integrative approach to investigating cancer.

Thus, arguably, the cancer genome project has itself led to a shift away from treating mutations as causally central in cancer, toward a more integrative, multifactor theoretical ontology in cancer. This shift in the ontological centrality of genes follows the theme of many big data projects. For instance, Leonelli (2012) has argued in the context of model organism research that bioontologies – such as the gene ontologies developed based on shared functional roles of genes in model organisms – is "constantly modified depending on the state of research and the interests of their users" (Leonelli, 2012). Likewise, Strasser (2019) demonstrates how – going back to the early 20th century – data collection can transform the questions we ask, and the categories we treat as central, not simply "testing" theories, but reframing how we understand what we ought to treat as central, in virtue of the very challenges faced by those attempting to put together systematic and usable collections of data.

In this way, both Leonelli and Strasser challenge the notion that "big data" research is a new kind of science. Leonelli (2016) argues that the model organism

research she focuses on is not "data driven" research, so much as a "data-centric" research. Hypotheses are at work at every stage in the classification and analysis of data. While this is not "experimental" science in the sense typically understood, it carries similar implications for reform of our theories. Likewise also, the above historical reconstruction of some episodes in TCGA's process of inception and response to challenges illustrates how various presuppositions about the central causal role of mutations in cancer were developed and refined. It would be a mistake to say these were "hypotheses" per se about the likely number and role of driver genes in cancer. The view that each cancer was "driven" by a small handful of mutations to specific genes was not purely "hypothetical" in the open ended sense of, "mere guesses." Rather, they were the presuppositions of the entire project, the product of multiple lines of research that had been ongoing for decades.

The extent of heterogeneity among cancers, and the sheer number of driver genes identified, was – at least by the lights of many of the researchers' accounts – surprising. They did not take into consideration a further presupposition: namely, the variety of background rates of mutation. Taking this into account in turn required them to rethink their conception of a "driver" gene. But, the concept of a driver is still a concept in flux. What it means to "control" or "drive" cancer has arguably become less rather than more definitive as a result of TCGA, as it has become more clear how heterogeneous not only cancers, but also the roles of genes in cancer progression, and the causal pathways involved in cancer, all are. There is a growing awareness that genes have both direct and indirect effects, that there are epigenetic factors that play important roles in gene expression in cancer, and that there are interactive effects between both genes, and the tissue microenvironment, that change over time, and are highly context sensitive. Mutations might play different roles over time; what and how they "drive" cancer is not a simple matter.

Yet, perhaps they ought not to have been so surprised by the extent of heterogeneity, and the lack of easily assignable causal mappings between genes and cancer phenotype. Developmental biologists have long known that different types of cells in different tissues differentiate in different ways, at different rates. Moreover, they've long been aware that tissue architecture varies across tissues and organs (Bissell, et. al.,

2011). A skin or lung or breast cell, before maturing, will divide a different number of times, and some cells in some tissues have higher rates of turnover, and senesce sooner, some later (Tomasetti, et. al., 2017). Indeed, this is no accident, but a finely tuned product of evolution. Cell, tissue, and organ development and differentiation requires tight regulation. So, the fact that different types of cancer might have different mutation rates, types of mutation, and thus extent of heterogeneity should not have been all that surprising.

Moreover, that there would be subpopulations of coevolving lineages in a tumor – and thus more or less heterogeneity between cancers even of the same subtype – should not have been surprising. In the 1970s, and even earlier, Cairns, Nowell, Doll and Hill, and others, had proposed that cancer is an evolving population of cells, with population dynamics (see, e.g., Morange, 2012; Frank, 2007; Plutynski, 2013). Taking into account the unique features of tissue microenvironment of each cancer type and subtype, anyone familiar with the dynamics of evolution in populations would have expected to find a great deal of heterogeneity in cancer cell populations.

Researchers engaged in TCGA were blindsided by the extent and nature of heterogeneity discovered in cancer cells perhaps in part because the "product" oriented framing of the research emphasized identifying "driver" mutations, which presumably were more focused targets of "actionable" drugs or other forms of intervention. This encouraged black boxing the environment in which cancer cells of different types, in different tissues and organs, found themselves. Only by incorporating this information were they in a better position to identify those genes most likely to play a significant role in cancer progression, versus those that – for instance, in AML – were simply a byproduct of stem cell division in hematopoetic stem cells as we age. Perhaps similar causes were driving the number and rate of mutations in different cancer types and subtypes. The question that comparison across cancers raises is whether, and to what extent, similar processes are driving different amounts and rates of acquisition of mutations in different cancers.

What TCGA illustrates is not a "new" kind of science, per se, but a very old process of iterated refinement of research questions, and generation of novel and more fine-grained hypotheses. It also illustrates the practical challenge of drawing upon and

integrating information across disciplinary divides. Focusing exclusively on one temporal and spatial scale can lead one to fail to predict complications and confounding causes of new discoveries. At the same time, reductive approaches can be heuristically useful, in that they enable the discovery of novel ways of framing research, and new questions one could have not anticipated without the new data.

To be sure, this is no small matter of relevance only to the basic scientists. The ambitions of precision oncology – targeted therapies for the many patients still suffering – depend in large part on quality and validation of purported cancer biomarkers, or molecular features associated with specific cancer mutations. The translation of this basic research into applications that enable precision patient care raises yet further practical and empirical challenges that provides ample opportunity for critical reflection on both big data science, and interdisciplinary, translational research – from bench, to bedside – by future historians, philosophers and social scientists of biomedicine, some of which is already underway (cf. Chin-Yee, et. al., 2019; Laplane, 2017, Fagan, 2016; Green, et. al., 2021; Vogt, et. al., 2019).

Acknowledgements:

I owe a great debt to Alan Love and Chris Donohue for organizing this workshop, and to Gar Allan, Janella Baxter, Lucie Laplane and Alan Love for comments on the manuscript. I also owe an immense debt to the scientists who agreed to be interviewed for this project: Li Ding, Carolyn Hutter, Malachi Griffith, Obi Griffith, Adrian Lee, Ramachandran Govindan, Timothy Ley, and Jane Peterson.

Bibliography

Bissell, M. J., & Hines, W. C. (2011). “Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression.” *Nature medicine* **17**(3): 320.

Boem, F., & Ratti, E. (2016). “Towards a notion of intervention in big-data biology and molecular medicine.” *Philosophy of Molecular Medicine: Foundational Issues in Research and Practice*, Boniolo & Nathan, eds. (147-164).

Böttcher, J. P., Bonavita, E., Chakravarty, P., Blees, H., Cabeza-Cabrerizo, M., Sammicheli, S., ... & e Sousa, C. R. (2018). “NK cells stimulate recruitment of cDC1 into the tumor microenvironment promoting cancer immune control.” *Cell* **172**(5): 1022-1037.

Cancer Genome Atlas Network. (2012). “Comprehensive molecular portraits of human breast tumours.” *Nature* **490**(7418): 61.

Chin-Yee, B., Sadikovic, B., & Chin-Yee, I. H. (2019). “Genomic data in prognostic models—what is lost in translation? The case of deletion 17p and mutant TP53 in chronic lymphocytic leukaemia.” *British journal of haematology* **188**(5): 652-660.

Collins, F. S., & Barker, A. D. (2007). “Mapping the cancer genome.” *Scientific American* **296**(3): 50-57.

Dawson, M. A. (2017). “The cancer epigenome: Concepts, challenges, and therapeutic opportunities.” *Science* **355**(6330): 1147-1152.

De Visser, K. E., Eichten, A., & Coussens, L. M. (2006). “Paradoxical roles of the immune system during cancer development.” *Nature reviews cancer* **6**(1): 24-37.

Ding, L. (Oct. 10, 2018) Personal Interview, Washington University in St. Louis

Ding, L., M. H. Bailey, E. Porta-Pardo, V. Thorsson, A. Colaprico, D. Bertrand, D. L. Gibbs et al. (2018) "Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics." *Cell* **173**(2): 305-320.

Esteller, M. (2008) “Epigenetics in cancer.” *N. Engl. J. Med.* **358**: 1148–1159.

Fagan, M. B. (2016). “Pathways to the clinic: cancer stem cells and challenges for translational research.” In *Philosophy of Molecular Medicine* (pp. 173-199). Boniolo & Nathan, Eds. Routledge.

Frank, S. A. (2007). *Dynamics of cancer: incidence, inheritance, and evolution*. Princeton: Princeton University Press.

Getz, G. 2015, (updated: 2016). “Challenges in Cancer Genomics” Broad Institute, MPG Primer. https://www.youtube.com/watch?v=2mIqL08uNRg&t=1992s

Govindan, R. July 6, 2018. Interview, Washington University in St. Louis.

Greaves, M., & Maley, C. C. (2012). “Clonal evolution in cancer.” *Nature* **481**(7381): 306.

Green, S., Dam, M. S., & Svendsen, M. N. (2021). “Mouse avatars of human cancers: the temporality of translation in precision oncology.” *History and Philosophy of the Life Sciences* **43**(1): 1-22.

Griffith, O. and M. April 2018. Personal Interviews, Washington University in St. Louis.

Griffith, O. L., Krysiak, K., Campbell, K., Spies, N., Kunisaki, J., Trani, L., ... & Griffith, M. (2019). “Surveying the genomic landscape of tumours and tumour models–the next frontier.” *Pathology* **51**: S4.

Gyanchandani, R., Y. Lin, H. Lin, K. Cooper, et. al., (2016) “Intra-tumor heterogeneity affects gene expression profile test prognostic risk stratification in early breast cancer.: *Clinical Cancer Research* **22**(21): 5362-5369.

Hey, S. P., & Barsanti-Innes, B. (2016). “Epistemology, ethics, and progress in precision medicine.” *Perspectives in biology and medicine* **59**(3): 293-310.

Hey, T., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Redmond, WA: Microsoft research.

Hoadley, K. A., C. Yau, T. Hinoue, D. M. Wolf, A. J. Lazar, E. Drill, R. Shen et al. (2018) "Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer." *Cell*. **173**(2): 291-304.

Hoadley, K. A., C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. DM Leiserson et al. (2014) "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin." *Cell* **158**(4): 929-944.

Hutter, C. July 27, 2018. Personal Interview, NHGRI, Washington D. C.

Jones, P. A., & Laird, P. W. (1999). “Cancer-epigenetics comes of age.” *Nature genetics* **21**(2): 163-167.

Keating, P., & Cambrosio, A. (2012). “Too many numbers: Microarrays in clinical cancer research.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43(1).

Kell, D. B., & Oliver, S. G. (2004). “Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era.” *Bioessays* **26**(1): 99-105.

Kimmelman, J. & AJ London (2015) “The Structure of Clinical Translation: Efficiency, Information and Ethics.” *Hastings Center Report* **45** (2): 27-39

Kimmelman, J., & Tannock, I. (2018). “The paradox of precision medicine.” *Nature Reviews Clinical Oncology* **15**(6): 341-342.

Laplane, L. (2017). *Cancer stem cells*. Harvard University Press.

Lawrence, M. S., P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter et al. (2013) "Mutational heterogeneity in cancer and the search for new cancerassociated genes." *Nature* **499**(7457): 214.

Lee, A. (2019). Personal Interview. University of Pittsburgh.

Leonelli, S. (2012). “Classificatory theory in data-intensive science: The case of open biomedical ontologies.” *International Studies in the Philosophy of Science* **26**(1): 47-65.

Leonelli, S. (2016). *Data-Centric Biology, A Philosophical Study*. Chicago: University of Chicago Press.

Ley, T. May 30, & July 18, (2018). Personal Interview, Washington University in St. Louis.

London, A. & Kimmelman, J. (2019). “Clinical Trial Portfolios: A Critical Oversight in Research Ethics, Drug Regulation and Policy.” *Hastings Center Report* **49**(4): 31-41.

Malta, T. M., A. Sokolov, A. J. Gentles, T. Burzykowski, L. Poisson, J. N. Weinstein, B. Kamińska et al. (2018). “Machine learning identifies stemness features associated with oncogenic dedifferentiation.” *Cell* **173**(2): 338-354.

Meyerson, M., Gabriel, S., & Getz, G. (2010). “Advances in understanding cancer genomes through second-generation sequencing.” N*ature Reviews Genetics* **11**(10): 685.

Miller DG. (1980). “On the nature of susceptibility to cancer.” The presidential address. *Cancer* **46**:1307–1318.

Morange, M. (2012). “What history tells us XXVIII. What is really new in the current evolutionary theory of cancer?” *Journal of biosciences* **37**(4): 609-612.

Nowak, M. and Waclaw, B. (2017) “Genes, Environment, and Bad Luck: Explaining Cancer Risk in the Statistical Sense.” *Science*. **355**: 6331.

Peterson, Jane. 2018. Personal communications. NHGRI: Workshop on History of Genomics, Organized by Alan Love and Chris Donohue

Plutynski, A. (2013). “Cancer and the goals of integration.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **44**(4): 466-476.

Plutynski, A. (forthcoming) “Why precision medicine is not very precise (and why we should not be surprised).” (editors, Marta Bertolaso, Chiara Beneduce) Personalized medicine. Springer.

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

Schinzel AC, Hahn WC. (2008) “Oncogenic transformation and experimental models of human cancer.” *Front. Biosci* **13**:71–84.

Shi, W., Ng, C. K., Lim, R. S., Jiang, T., Kumar, S., Li, X., ... & Weigelt, B. (2018). “Reliability of whole-exome sequencing for assessing intratumor genetic heterogeneity.” *Cell reports* **25**(6): 1446-1457.

Strasser, B. J. (2012). “Data-driven sciences: From wonder cabinets to electronic databases.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **43**(1): 85-87.

Strasser, B. (2019) *Collecting Experiments: Making Big Biology*. Chicago: University of Chicago Press.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). “The cancer genome.” *Nature* **458**(7239): 719.

Tannock, I. F., & Hickman, J. A. (2017). “Limits to precision cancer medicine.” *The New England journal of medicine* **376**(1): 96.

Tao, D & V. Prasad (2018) “Choice of control group in randomised trials of cancer medicine: are we testing trivialities?” *The Lancet Oncology* **19**: 1150-1152.

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Yang, T. H. O., ... & Chuah, E. (2018). “The immune landscape of cancer.” *Immunity* **48**(4): 812-830.

Smalheiser, N. R. (2002). “Informatics and hypothesis-driven research.” *EMBO reports* **3**(8): 702-702.

Tomasetti C. and B. Vogelstein (2015) “Variation in cancer risk among tissues can be explained by the number of stem cell divisions.” *Science* **347**: 78

Tomasetti, C., L. Li, B. Vogelstein, (2017) “Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention.” *Science* **355**: 1330.

Vander Heiden, M. G., & DeBerardinis, R. J. (2017). “Understanding the intersections between metabolism and cancer biology.” *Cell* **168**(4): 657-669.

Vogt H, Green S, Ekstrøm CT, Brodersen J. (2019) “How precision medicine and screening with big data could increase overdiagnosis.” *BMJ*. **366**: l5270.

Welch, J. S., Ley, T. J., Link, D. C., Miller, C. A., Larson, D. E., Koboldt, D. C., ... & Kandoth, C. (2012). “The origin and evolution of mutations in acute myeloid leukemia.” *Cell* **150**(2), 264-278.

Wu, C., S. Powers, W. Zhu, Y. Hannun, (2016) “Substantial contributions of extrinsic risk factors to cancer development.” *Nature 529*: 43.