HHS Public Access

Author manuscript

Clin Cancer Res. Author manuscript; available in PMC 2023 February 15.

Published in final edited form as:

Clin Cancer Res. 2022 August 15; 28(16): 3417–3424. doi:10.1158/1078-0432.CCR-19-3748.

Data-rich spatial profiling of cancer tissue: Astronomy informs Pathology

Alexander S. Szalay, PhD[1,4,5], **Janis M. Taube, MD**[2,3,4,*]

[1]Dept of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA

[2]Depts of Dermatology, Pathology, and Oncology, Johns Hopkins University SOM, Baltimore, MD 21287, USA

[3]The Bloomberg~Kimmel Institute for Cancer Immunotherapy at Johns Hopkins University

[4]The Mark Foundation Center for Advanced Genomics and Imaging at Johns Hopkins University

[5]Dept of Computer Science, Johns Hopkins University

Abstract

Astronomy was among the first disciplines to embrace Big Data and use it to characterize spatial relationships between stars and galaxies. Today, medicine, in particular pathology, has similar needs with regard to characterizing the spatial relationships between cells, with an emphasis on understanding the organization of the tumor microenvironment. In this paper, we chronicle the emergence of data-intensive science through the development of the Sloan Digital Sky Survey and describe how analysis patterns and approaches similarly apply to multiplex immunofluorescence (mIF) pathology image exploration. The lessons learned from astronomy are detailed, and the new AstroPath platform that capitalizes on these learnings is described. AstroPath is being used to generate and display tumor-immune maps that can be used for mIF immuno-oncology biomarker development. The development of AstroPath as an open resource for visualizing and analyzing large scale spatially-resolved mIF datasets is underway, akin to how publicly-available maps of the sky have been used by astronomers and citizen scientists alike. Associated technical, academic, and funding considerations, as well as extended future development for inclusion of spatial transcriptomics and application of artificial intelligence, are also addressed.

Keywords

AstroPath; PD-1; pathology; astronomy; big data; biomarker

[*]To whom correspondence should be addressed: Janis M. Taube, MD, Director of Dermatopathology Division, Blalock 907, 600 N. Wolfe St., Baltimore, MD 21287, Tel: 410-955-3484, jtaube1@jhmi.edu.

COI: Dr. Taube reports grants and consulting from Bristol-Myers Squibb, consulting for Merck, Astra Zeneca, and Compugen outside the submitted work; Drs. Taube and Szalay report equipment and reagents as well as stock options from Akoya Biosciences. Drs. Taube and Szalay also have a patent pending related to image processing of mIF/IHC images.

INTRODUCTION

Throughout history, people have looked to the stars for answers to vital questions. Until the 18th century, astrology formed the bridge between Astronomy and Medicine, and the study of the motions of celestial bodies were often linked the rhythms of the human body or specific organ systems, Figure 1.[1] Vestiges of this relationship remain even today, e.g. the term 'lunatic', which derives from the concept that the moon's trajectory impacted human physiology. There are also a number of exceptional medical astronomers. For example, while Nicolas Copernicus is best known for pioneering the heliocentric theory, he was a practicing medical doctor. Three centuries later in 1839, John William Draper, MD, took the first photograph of the moon.[1]

DATA-INTENSIVE SCIENCE: THE FOURTH PARADIGM

Both medicine and astronomy have become pure sciences over the last few centuries, and the current link between these two disciplines is more tenuous than in earlier centuries. However, the relationship is converging again, due to the emergence of Big Data and the Fourth paradigm of science. Jim Gray (in collaborating with Szalay) called this the Fourth Paradigm of science .[2,3] In brief, thousands of years ago, science was largely empirical in nature and focused on the description of natural phenomena, *e.g.,* the Chinese star charts. Then hundreds of years ago, theoretical science emerged as the Second Paradigm, expressing physical laws as mathematical models, abstractions *e.g.,* Kepler's or Newton's Laws, etc., and was exemplified by the work of Albert Einstein. Advanced computing ushered in the Third Paradigm, as it allowed for the computational simulation of complex phenomena such as atomic or molecular dynamics. Approximately 20 years ago, the Fourth Paradigm of (Big) data-driven science started to emerge. Astronomy has been the flagship discipline for this Fourth Paradigm, in its uniform approach to data that transcends scientific discipline. Its key pillars include data mining, statistical learning, pattern detection, and artificial intelligence.

This shift towards data-intensive science has implications for experimental design and associated workflows. The conventional approach has been that data is first acquired, then analyzed, and the results are published. Additional confirmatory studies may then be performed after publication. In contrast, Big Data-based studies revert this paradigm, Figure 2. Here, data is acquired, calibrated and organized, essentially 'published' before data analysis is performed, correlations are discovered and then confirmatory studies performed. Put another way, the raw data is processed or presented, a.k.a. 'published' in such a way that it becomes fodder for exploration and analysis, and in many instances the hypotheses to be explored are often derived from the data itself. Astronomy was a natural leader in this space, as the ability to conduct true laboratory experiments is prohibitive, and instead passive observations of data are typically required. In this model, scientists are forced to become publishers. Data may be organized into a database, which can then be opened up to everyone. As such, the scientists serving the data become trusted intermediaries to external users, and this often requires lots of specialized expertise in handling and processing raw data and organizing it into an accessible database.

SLOAN DIGITAL SKY SURVEY (SDSS)

Before SDSS was developed, astronomers performed bespoke analyses in their own institutions using customized processing algorithms. The vast majority of an astronomers' time was spent generating the data, and in the end, it was often difficult to directly compare scientific analyses across laboratories, given the site-to-site variations in data acquisition and handling. SDSS I-II was a 16-year effort (1992–2018) that transcended this issue.[4,5] SDSS used a dedicated telescope in New Mexico, which acquired approximately 200 GB of images of the sky each night. Raw patches/tiles of sky were processed in an overlapping fashion that allowed for correction of various instrumental artifacts as well as facilitated stitching of a seamless and accurate map of the sky. The raw image tiles were automatically processed and incorporated into a spatial database. It opened to the public in 2001, and provided annual updates ever since (SDSS-V has just started). In the end, the SkyServer hosted several hundred terabytes of multispectral imaging and spectroscopy data that were collected and processed in a uniform fashion.

SDSS transformed astronomy. With high-quality data provided, astronomers were able spend the majority of their time on building creative scientific hypotheses via data exploration, rather than generating the primary data itself. To date, over 9,000 refereed publications and 500,000 citations have been based on SDSS data. The open platform of SDSS, the SkyServer, allowed not just astronomers, but citizen scientists to access the publicly-served data, leading to important discoveries by non-astronomers including a Dutch school teacher and a group of computer gamers.[6,7]

Importantly, the SkyServer contained a rich and evolving toolset of spatial functions and analytics, integrated with a relational database into a collaborative science platform[8]. Its most important feature was its "shrink-wrapped" quality: providing a data set that was statistically as uniform as possible, properly calibrated against systematic errors, well documented, had information about the lineage and provenance of the data and all processing steps, provided deep links to the underlying raw data, and was presented through tools which are intuitive and easy to use. Furthermore, the data was stable over extended periods, allowing time to build trust with the user community.

THE ASTROPATH PLATFORM

The SkyServer has evolved into the SciServer,[8] hosting data from a broad array of disciplines through the Institute for Data Intensive Engineering and Sciences (IDIES; http://idies.jhu.edu/) at the Johns Hopkins University. Projects are as diverse as studying turbulence,[9,10] ocean circulation models,[11] soil ecology, and digital sociology,[12,13] including mapping abandoned row houses in Baltimore to help determine which ones should be rebuilt rather than torn down to provide more greenspace in neighborhoods. Medical applications include supporting genomic datasets[14,15] and mapping neural connectomes.[16]

Most recently, we have begun to apply our experience in hosting and querying spatially-resolved data to tissue pathology images. The specific use case that initiated this collaboration was the desire to quantify multispectral, multiplex immunofluorescence (mIF)

images of tumor tissue across large cohorts of patients. The next generation of prognostic and predictive biomarkers will capitalize on the spatial arrangements of cells *in situ*, and their distinct co-expression profiles. One particular area of interest is studying tumor tissue in the context of immuno-oncology, as the interaction of the immune system with tumors has prognostic implications.[17–23] Spatial representations of this interplay can also be used as biomarkers for immunotherapy.[24–26] Emerging platforms that use multiplex immunofluorescence (mIF) and spatial transcriptomic approaches have the potential to characterize the complexity of the tumor-immune interaction more completely than ever before, and it is now possible to label individual cells with 5–50 markers in the case of protein labels and regions of interest with hundreds of transcripts in the case of mRNA expression profiling.

The long-term goal of the AstroPath effort is to generate spatial maps of tissue using these multiple modalities, such that the data generated is both large in amount in the tissue covered and high-dimensional in the number of markers mapped. The amount of tissue mapped vs. the number of markers is a compromise, and different investigative groups and platform developers value this trade-off differently. The key then to multimodal integration of tissue imaging is to have a foundational, reference map with a broad and deep enough feature vector to be able to cover the whole tissue. mIF with ~5–8 markers is a robust method that can be used to serve this purpose, and the resultant mIF maps can then form the scaffolding by which other spatial-immune data from other platforms, *e.g.* spatial transcriptomics, or even higher order mIF, etc, can be further contextualized. There are continued analogues to astronomical mapping. In fact, the current status of spatial transcriptomics (high dimensionality in number of markers tested, but with less resolution and smaller area covered) is analogous to how spectroscopic data was layered onto discrete object imagery in the SDSS (the MANGA instrument with integral-field spectroscopy). Similarly, the Hubble data was very deep but only covered ~45 square degrees of 40,000 in the whole sky, while the SDSS covered 15,000 square degrees but in much less depth. This is also akin to how the higher-plex approaches to mIF –such as tissue mass cytometry imaging with ~40–50 markers – may be achieved, yet are currently accompanied by the trade off of less surface area covered.

There are many additional direct comparisons between mIF tissue imaging and where the astronomy field was a few decades ago, such as the need to scale and organize multicolor photometric data, acquire and segment images, optimize signal to noise ratios for low-level signal detection, obtain robust measurements of marker expression intensity, and perform spatial statistics, Figure 3. To begin to address these challenges, we developed a mIF workflow that included standards for staining, image acquisition and processing, segmentation and associated phenotyping of cells, controlling for batch effects, and data handling, Figure 4. Importantly, rather than imaging singular high-power fields (HPFs) on a slide, we adopted the approach used by SDSS, whereby overlapping HPF image tiles were acquired across the whole slide, Figure 5. The image tiles overlapped the neighboring tiles by 20% surface area, facilitating correction of inherent, residual lens-distortions, flat-fielding of microscope optics and seamless stitching of whole slide images and an *a priori* selection of a statistically fair subset of the image pixels for further analyses. The end result was

highly accurate tumor-immune maps with single cell resolution, organized in a relational database that could be queried similar to the SkyServer.

We used the tumor-immune maps and developed an approach for reducing whole-slide imagery to an actionable, biomarker score for predicting response and long-term outcomes following anti-PD-1-based therapy for patients with melanoma.[27] Key features measured as a part of this score included the intensity of PD-1 expression, the identification of rare cell types such as CD8+FoxP3+cells and tumor or CD163+ cells that were PD-L1 negative. This required the accurate quantification of PD-1 and PD-L1 intensity *in situ*, which was achieved in the AstroPath database with corrections for illumination variation and a standardized approach to batch-to-batch correction. The unbiased acquisition of the whole slide in a tiled fashion also allowed for testing various slide-sampling strategies in a rational, user-independent manner. We showed that the predictive value for detecting the rare CD8+FoxP3+ population was improved when ~10–30% of the slide with the highest CD8 densities was sampled. Such comparisons had not previously been performed for mIF assays and are important for biomarker optimization and standardization prior to clinical implementation.

We are now expanding the number of pre- and on-treatment tumor types mapped using the AstroPath platform, and currently have over half a billion individual cells spatially mapped across the tumor microenvironment (TME) in the database. Our long-term goal is to build the pathology equivalent of the SkyServer, including an interactive AstroPath browser with a navigation interface that can zoom seamlessly from whole-slide images of tumor tissue to individual HPFs and individual cells with a few mouse-clicks. The viewer will flexibly display boundary overlays and centroids of individual cells, save them in an online “lab-notebook” and export lists of marked regions. The tool will also enable pathologists to create custom annotations and store them in their own part of the database. Most of the necessary computations will be performed server-side, thus requiring only a standard light-weight web browser for the end-user conducting the analysis. We plan to build custom aggregation functions that can feed into various higher-level (R/Python) statistical packages, running in Jupyter notebooks. These tools will include distribution functions or characterizations of the intercellular distances of various subsets and will be both staining panel and platform agnostic, i.*e.,* can host data from MIBI, Codex, Vectra, and other such technologies including those focused on spatial transcriptomics. This will make scalable analyses involving tens to hundreds of millions of cells as easy as working with a small Excel spreadsheet today.

LESSONS LEARNED

There are several lessons from astronomy, learned the hard way, which were laid out in a systematic fashion before development of the AstroPath platform began. The main design principles are listed below:

a. Scalability is hard. Maximal automation and parallelism should be built in from the very beginning. Data should flow seamlessly from instrument to database, with minimal (possibly with zero) human intervention. This requires a well-

defined and strict protocol on file formats, naming conventions and directory structure, laid out in a short reference document. Human effort can then more efficiently used for executive oversight.

b. The whole pipeline should be designed with extreme hardware parallelism in mind, so that in the future each slide can be processed on separate servers, if needed, up to the final database.

c. Having the final data in a well-designed relational database makes data consistency and reproducibility much easier. Self-documenting, consistent metadata must be fully part of the database from the beginning.

d. With large enough data sets statistical errors become small, and systematic errors dominate. These can only be discovered with a conscious systematic effort, using redundant observations. Statistical reproducibility requires a well-thought out quality assurance and calibration process. Error estimates must be derived objectively, internally, from multiple independent observations of the same physical pixels.

e. Everything is spatial. All geometric information, positions, and shapes should be captured and calibrated carefully, since in the studies of the TME, relative proximities are extremely important. Various spatial operations on points and polygons (buffer, intersection, union, search, distance) should be implemented inside the database for speed. It should be possible to locate objects (cells) within arbitrary geometric regions. A flexible visual browser capable of displaying all this information from the whole slide to the level of the individual cells is needed.

f. Cloud computing has a lot of appeal, but is best adopted at specific project stages in an elastic fashion. Long-term cloud-based storage is currently 7–10x more expensive than keeping servers locally. It is also 3x more expensive than locally-kept (and fully utilized) AI-computing infrastructure, but for sparse use the Cloud is already cheaper. The clear advantage to Cloud computing is apparent when large, transient computing capability is needed. In the long run, once Cloud pricing changes, this will inevitably become part of our computing infrastructure. Nonetheless, we'll always need significant storage immediately next to the microscopes to capture the immediate data.

REMAINING CHALLENGES

There are also a number of differences between characterizing astronomical objects and those of the tumor microenvironment, many of which also represent the ongoing challenges with multiplex immunofluorescence image handling. In SDSS, the vast majority of the data was collected using one main telescope, whereby with mIF, there is a large assortment of different platforms and reagents in use, and the proliferation of technologies is not expected to abate in the near future. Imaging of the sky also requires correction for conditions of the atmosphere, including weather and variations in atmosphere thickness. Astronomers have had to develop calibrations that allow them to calculate the brightness of the emissions

as if there were no atmosphere, while algorithms have yet to be able to correct for pre-analytic variation in tissue handling practices. Another difference is that in astronomy, many pixels are empty sky, with much smaller fraction covered by “useful” fluxes. As such, segmentation of images into individual cells (rather than stars or galaxies) is more challenging in the densely packed cellular environment of a tumor. These and associated challenges are described in detail below.

Technical

We have completed the first proof of principle experiments and mapped out the pipeline with AstroPath with regard to ‘manufacturing’ data. We would now like to undergo an ‘industrial revolution’ where mass production of data is our norm. The genomics community have undergone this transition, and pathology is now poised to cross over. We proposed a standardized approach for each step in the generation of mIF data, with an eye to modularity and parallelism.[17] In this model, data will be generated from multiple microscopes running simultaneously, and each will have to be calibrated individually to ensure they generate equivalent results. While this may be controlled within a single laboratory, numerous different platforms with different specifications and output formats are being used across the community, which is a major, unsolved challenge. Further, even within a given platform, the microscopes themselves and associated reagents are still in development and are changing over time. Numerous analytic pipelines are emerging and evolving as well.[27–29] For a true community archive of mIF images to be developed, such methodology will need to become more formalized and consistent across laboratories. In the meantime, best practices for reproducible image acquisition and analysis will need to be agreed upon, akin to what has been described for chromogenic, in situ hybridization tests, and mIF slide staining.[30–32] Image analysis pipelines that can ingest data from different platforms or which capitalize on pre-existing, standardized workflows (such as those for flow cytometry), are also an important step towards this goal.[29,33,34]

As extremely large data sets are collected, statistical noise will decrease, but it is anticipated that systematic errors will become more of a concern. Variation in instrument calibration is only one of the many possibilities that could fall into this category, and there are likely errors that will be experienced that we are not even aware of yet today. Artificial Intelligence should be very good at detecting hidden patterns in the data that may be correlated with instrumental sources, and alerting users to and even potentially correcting for these systematic differences. As datasets grow, Artificial Intelligence will be applied to the data sets, and will undoubtedly identify subtle patterns that are of biological import. Some exciting developments in this area include those that facilitate ‘interpretable’ weakly supervised deep learning.[35] These approaches highlight the features that the machine learning model uses to make the decision, thus potentially revealing previously unappreciated or underappreciated facets of the TME.

Another major hurdle that has yet to be addressed in a meaningful way is the impact of pre-analytic variables in large pathology datasets. Pre-analytic variables in tissue collection, processing, and storage can impact mIF staining performance. It will be important to define parameters such as sample age, and interfering substances such as melanin or anthracotic

pigment, which have the potential to impact marker signal acquisition. This has been a time-honored challenge in pathology, but is even more important if the goal is to measure the intensity of marker fluxes in situ, as opposed to just scoring staining as a binary positive vs. negative.[27] Once these factors are defined, it may be possible to prospectively collect and keep specimens according to the requisite conditions. It would also be advantageous to be able to correct for pre-analytic variables in specimens that have already been collected under diverse conditions, akin to how the astronomers have addressed variation in atmospheric conditions. This will require a sophisticated calibration, as the different degrees of freedom in tissue samples are greater, yet it is worth pursuing—such calibrations would ensure that mIF studies could be reproducibly conducted across pre-existing archives of formalin-fixed paraffin embedded tissues from surgical pathology laboratories from around the world.

Academic and funding considerations

Academia offers a fertile environment for large multidisciplinary and "antedisciplinary"[36] efforts such as AstroPath, which includes experts in Big Data, pathology, immunology, oncology, engineering, computer vision, biostatistics, etc. However, the attribution of academic credit for such efforts can be challenging using the traditional currency of paper publications. The creation of large datasets and associated pipelines to generate them are their own area of specialized expertise, yet there are several questions surrounding this new publication form – including whether they have to be 'published' through traditional publishing houses, how to recognize authors of the dataset itself, and whether they need to be peer-reviewed. Often when large datasets are published, it is as supplementary materials to a more traditional research publication, which may not provide appropriate credit to those who generated the foundational dataset. Astronomy has addressed this by partnering with a print journal and releasing descriptions yearly about updates to the dataset and providing a new link to the latest dataset version in a manuscript format.[37] Each year, the authors generating the latest version of the dataset are represented on that print publication, which is the official academic release of the data, yet independent of any scientific studies that may have been conducted on the dataset. The mechanism by which peer-review of such published datasets would be conducted remains an open question, as review of each SDSS TB-sized data release has not been practicable. One can argue that the transparency, consistency, and stability surrounding data releases over the two decades, along with the ~9,000 refereed publications to date, indicate the trust the astronomy community has in the SDSS data, and provides a reputation-based metric. More than half of these papers are by non-SDSS authors. In astronomy, a typical paper costs ~$100,000 in students/post-doc time and computing hardware. Thus, 5,000 papers represent close to $500M of research dollars from non-SDSS-supported resources that have been committed over the last 20 years to explore this dataset, representing another relevant metric of its immense impact and associated user confidence.

Funding for generating such datasets deserves specific consideration. The Sloan Foundation largely funded SDSS (augmented by National Science Foundation, Department of Energy, and NASA), and the continued AstroPath platform development is currently funded near-completely by private foundations, many of which have connections to entrepreneurial, billionaire investors. The phases of a project's lifetime include its architecture, prototyping, implementation, and then sustainability. Start-ups often fail at the transition between

prototyping and a fuller implementation. This “Valley of Death” is the critical time when a polished product needs to be presented to users/customers to get them to abandon their current approach and move over to the new solution. Supporting this key transition after a prototype makes it to a beta-version is difficult in the traditional peer-reviewed setting. Probably the best hope is in public/private partnerships, where riskiest parts of developing a new breakthrough dataset are underwritten by a private foundation, while the longer-term operation is supported by a federal grant.

Critical, well calibrated data sets have an increasingly important role for Artificial Intelligence. Many of the recent breakthroughs in Deep Learning came from the combination of such a unique data set with a disruptive idea.[38] Every AI project needs large, well labelled training sets containing ground truth data. The ultimate quality of the resulting inferences from Deep Learning will reflect upon the quality of the input data. As a result, the community will need much bigger and better public data sets which are “AI-ready”. We hope that the approach outlined here will help to convince our peers to join a wide community-based effort to share well-calibrated, high-quality data through an open cancer cell repository, what will support a variety of cross-cutting explorations.

ACKNOWLEDGEMENTS

This work was supported by The Mark Foundation for Cancer Research (JMT, AS); Melanoma Research Alliance (JMT, AS); Moving for Melanoma of Delaware (JMT); Sidney Kimmel Cancer Center Core Grant P30 CA006973 (JMT); National Cancer Institute R01 CA142779 (JMT,AS); and The Bloomberg~Kimmel Institute for Cancer Immunotherapy.

REFERENCES

1). Strach EH. Astronomy and Medicine. J Brit Astron Assoc 1982;92:164–9.

2). Hey AJG. The Fourth Paradigm: data-intensive scientific discovery Microsoft Research Publishing. Redmond, WA. 2009.

3). Szalay AS and Gray J. Science in an Exponential World. Nature 2006;440:23–4. [PubMed: 16511466]

4). Szalay AS and Gray J: The World Wide Telescope. Science 2004;293:2037–40.

5). Raddick MJ, Thakar AR, Szalay AS, Santos RDC. Ten Years of SkyServer I: Tracking Web and SQL e-Science Usage. Computing in Science and Engineering 2014;16:22–31.

6). Lintott CJ, Schawinski K, Keel W, van Arkel H, Bennert N, Edmondson E, et al. Galaxy Zoo: ‘Hanny’s Voorwerp’, a quasar light echo? Monthly Notices of the Royal Astronomical Society 2009;399:129–40.

7). Cardamone C, Schawinski K, Sarzi M, Bamford SP, Bennert N, Urry CM, et al. Galaxy Zoo Green Peas: discovery of a class of compact extremely star-forming galaxies. Monthly Notices of the Royal Astronomical Society 2009;399:1191–205.

8). Szalay AS. From SkyServer to SciServer. The Annals of the American Academy of Political and Social Science 2018;675:202–20.

9). Li Y, Perlman E, Wan M, Yang Y, Charles Meneveau C, Burns R, et al. A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence. Journal of Turbulence 2008;9:1–29.

10). The JHU Turbulence database: http://turbulence.pha.jhu.edu

11). Almansi M, Gelderloos R, Haine T, Saberi A, and Siddiqui A. Oceanspy: A python package to facilitate ocean model data analysis and visualization. The Journal of Open Source Software 2019;4:1506.

12). Garboden PME, Fan L, Budavari T, Basu A, Evans JD. Combinatorial Optimization for Urban Planning: Strategic Demolition of Abandoned Houses in Baltimore Working Papers, 2019–5, University of Hawaii Economic Research Organization, University of Hawaii at Manoa. (2019).

13). Lapowski I. The Astrophysicist Who Wants to Help Solve Baltimore's Urban Blight. Wired; 2018. https://www.wired.com/story/baltimore-vacant-houses-astrophysicist-algorithm/

14). Wilton R, Wheelan SJ, Szalay AS, Salzberg SL. The Terabase Search Engine: A Large-Scale database of short-read sequences. BioInformatics 2019;35:665–70. [PubMed: 30052772]

15). Wilton R, Szalay AS. Arioc. High-concurrency short-read alignment on multiple GPUs. PLoS computational biology;2020:16(11),e1008383. [PubMed: 33166275]

16). Burns R, Vogelstein JT, Szalay AS. From Cosmos to Connectomes: The Evolution of Data-Intensive Science. Neuron 2014;83:1249–52. [PubMed: 25233306]

17). Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science 2006;313:1960–4. [PubMed: 17008531]

18). Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer 2012;12:298–306. [PubMed: 22419253]

19). Feng Z, Bethmann D, Kappler M, Ballesteros-Merino C, Eckert A, Bell RB, et al. Multiparametric immune profiling in HPV– oral squamous cell cancer. JCI Insight 2017;2(14):e93652.

20). Gartrell RD, Marks DK, Hart TD, Li G, Davari DR, Wu A, et al. Quantitative Analysis of Immune Infiltrates in Primary Melanoma. Cancer Immunol Res 2018;6:481–93. [PubMed: 29467127]

21). Jackson HW, Fischer JR, Zanotelli VRT, Ali HR, Mechera R, Soysal SD, et al. The single-cell pathology landscape of breast cancer. Nature 2020;578:615–20. [PubMed: 31959985]

22). Schürch CM, Bhate SS, Barlow GL, Phillips DJ, Noti L, Zlobec I, et al. Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. Cell 2020;182:1341–59. [PubMed: 32763154]

23). Taube JM, Anders RA, Young GD, Xu H, Sharma R, McMiller TL, et al. Colocalization of inflammatory response with B7-h1 [PD-L1] expression in human melanocytic lesions supports an adaptive resistance mechanism of immune escape. Sci Transl Med 2012;4:127ra37.

24). Giraldo NA, Nguyen P, Engle EL, Kaunitz GJ, Cottrell TR, Berry S, et al. Multidimensional, quantitative assessment of PD-1/PD-L1 expression in patients with Merkel cell carcinoma and association with response to pembrolizumab. J Immunother Cancer 2018;6(1):99. [PubMed: 30285852]

25). Topalian SL, Bhatia S, Amin A, Kudchadkar RR, Sharfman WH, Lebbé C, et al. Neoadjuvant Nivolumab for Patients With Resectable Merkel Cell Carcinoma in the CheckMate 358 Trial. J Clin Oncol 2020;38:2476–87. [PubMed: 32324435]

26). Johnson DB, Bordeaux J, Kim JY, Vaupel C, Rimm DL, Ho TH, Quantitative Spatial Profiling of PD-1/PD-L1 Interaction and HLA-DR/IDO-1 Predicts Improved Outcomes of Anti–PD-1 Therapies in Metastatic Melanoma. Clin Cancer Res 2018;24:5250–60. [PubMed: 30021908]

27). Berry S, Giraldo NA, Green BF, Cottrell TR, Stein JE, Engle EL, et al. Analysis of multispectral imaging with the AstroPath platform informs efficacy of PD-1 blockade. Science 2021;372(6547);eaba2609. [PubMed: 34112666]

28). Greenwald NF, Miller G, Moen E, Kong A, Kagel A, Dougherty T, et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. Nat Biotechnol 2021 Nov 18. doi: 10.1038/s41587-021-01094-0. Online ahead of print.

29). Schapiro D, Sokolov A, Yapp C, Chen YA, Muhlich JL, Hess J, et al. MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging. Nat Methods 2021 Nov 25. doi: 10.1038/s41592-021-01308-y. Online ahead of print.

30). Taube JM, Akturk G, Angelo M, Engle EL, Gnjatic S, Greenbaum S, et al. The Society for Immunotherapy of Cancer statement on best practices for multiplex immunohistochemistry (IHC) and immunofluorescence (IF) staining and validation. J Immunother Cancer 2020;8(1):e000155. [PubMed: 32414858]

31). Taube JM, Roman K, Engle EL, Wang C, Ballesteros-Merino C, Jensen SM, et al. Multi-institutional TSA-amplified Multiplexed Immunofluorescence Reproducibility Evaluation (MITRE) Study. J Immunother Cancer 2021;9:e002197. [PubMed: 34266881]

32). https://www.nordiqc.org/

33). Schapiro D, Jackson HW, Raghuraman S, Fischer JR, Zanotelli VRT, Schulz D, et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. Nat Methods 2017;14:873–6. [PubMed: 28783155]

34). Giraldo NA, Berry S, Becht E, Ates D, Schenk KM, Engle EL, et al. Spatial UMAP and Image Cytometry for Topographic Immuno-oncology Biomarker Discovery. Cancer Immunol Res 2021;9:1262–9. [PubMed: 34433588]

35). Jiménez-Sánchez D, Ariz M, Chang H, Matias-Guiu X, de Andrea CE, Ortiz-de-Solórzano C. NaroNet: Discovery of tumor microenvironment elements from highly multiplexed images. Med Image Anal 2022;78:102384. [PubMed: 35217454]

36). Eddy SR. ‘‘Antedisciplinary’’ Science. PLOS Comput Biol 2005;1:e6. [PubMed: 16103907]

37). https://www.sdss.org/science/data-release-publications/

38). Krizhevsky A, Sutskever I and Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks, in Advances in Neural Information Processing Systems NIPS2012, eds Pereira F and Burges CJC and Bottou L and Weinberger KQ, Curran Associates, Inc. c399862d, Vol 25, (2012).

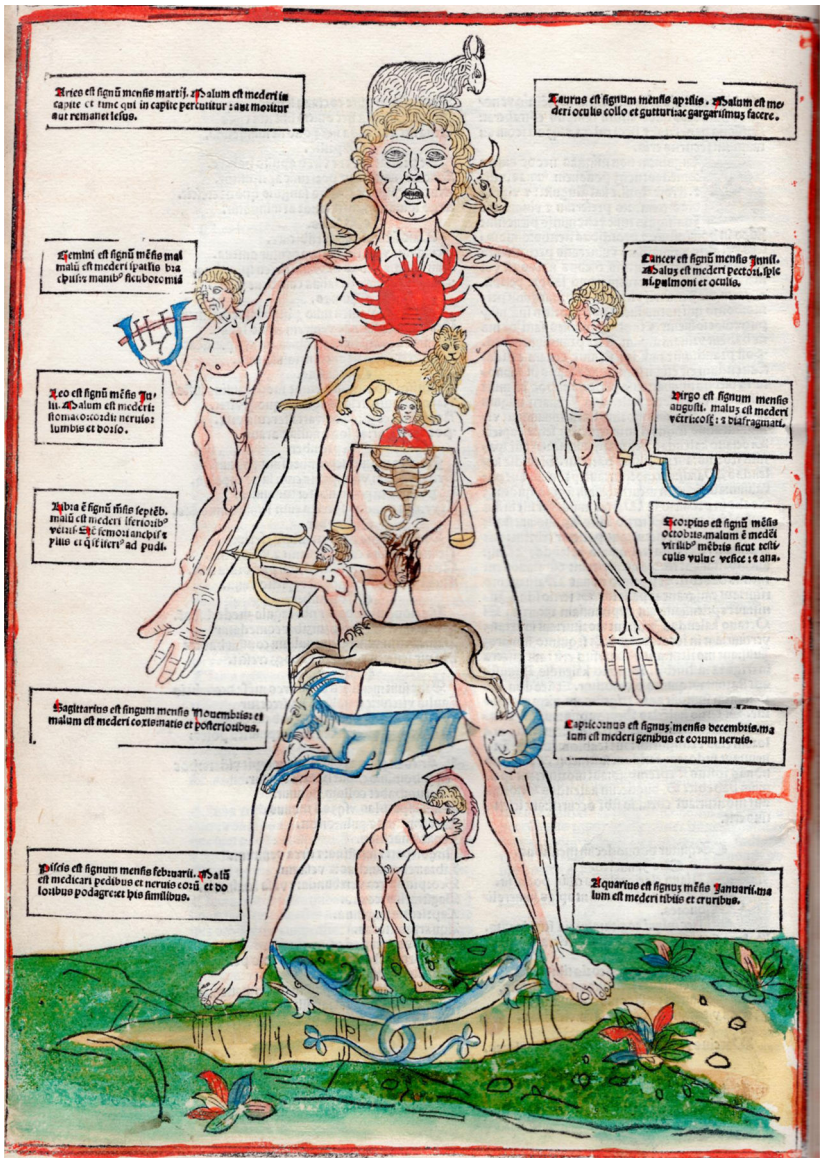

Figure 1. 'Astrology Man' relates the constellations to the human body.

For example, Leo is related to the heart, and Capricorn is related to the knees (Ketham, Joannes de, et al. *Fasciculus Medicinae.* Venetiis: Joannem and Gregorium de Gregoriis, fratres, 1500. Historical Collection, Institute of the History of Medicine, The Johns Hopkins University; used with permission). The constellations were the first attempt at a spatial index on a celestial sphere, illustrating an early understanding of the need for such a reference system in the history of medicine.

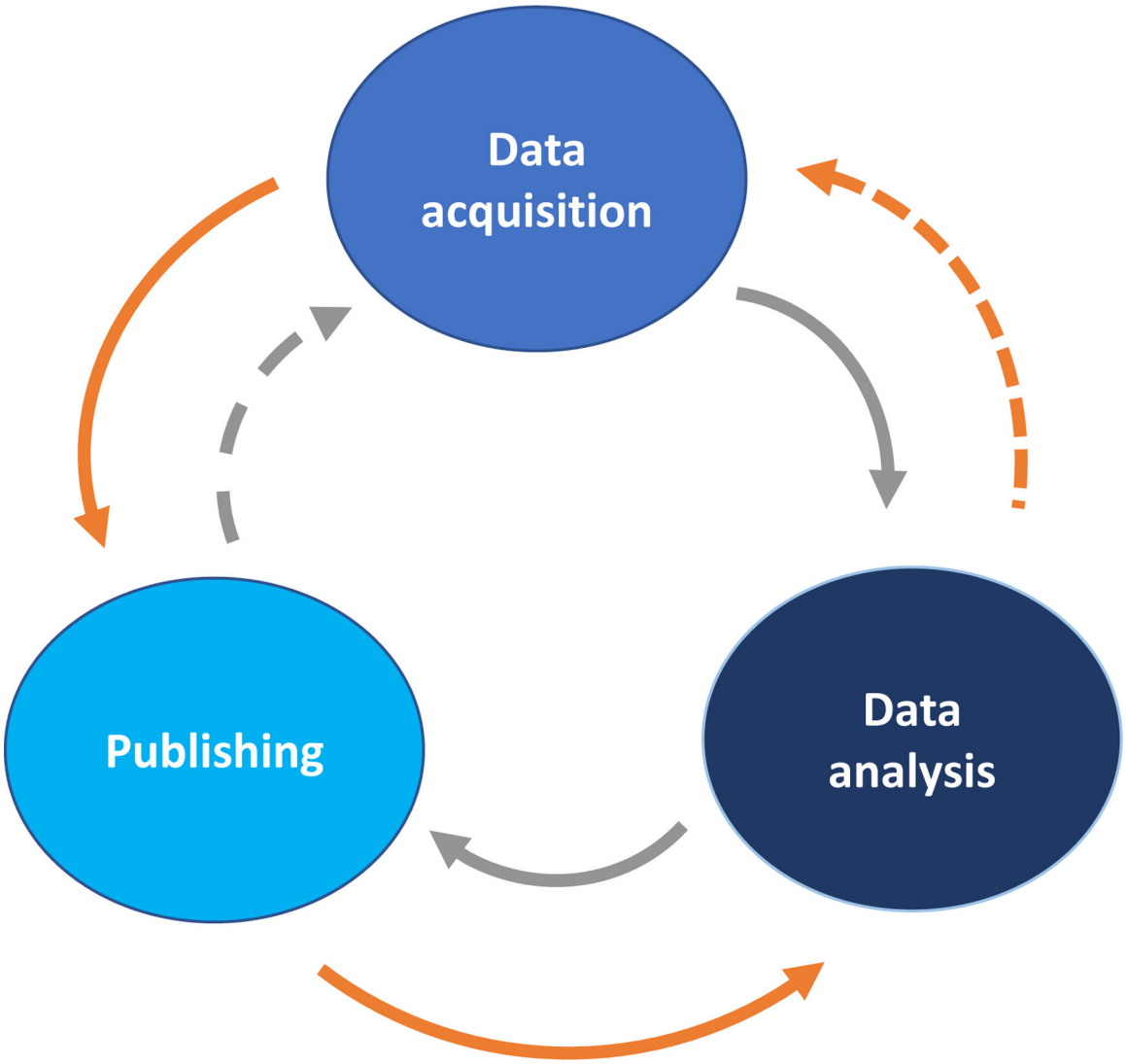

Figure 2.
The time-honored relationship between data acquisition, analysis, and publishing (grey) is reversed in the era of Big Data (orange).

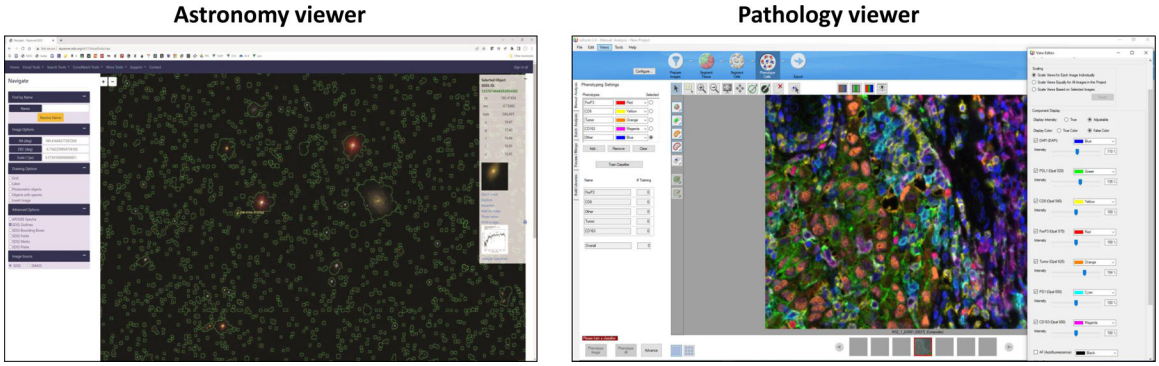

Figure 3. Cells in pathology are similar to stars and galaxies in astronomy.

The astronomy viewer for the SkyServer is shown on the left, and a pathology viewer for multispectral, mIF is shown on the right. Similar functionalities in both viewers include the ability to assess spectral strength, segment the image into discrete objects, and quantify spatial relationships between objects, amongst others. Images courtesy of SDSS (the Sloan Digital Sky Survey; www.sdss.org) and Akoya Biosciences, respectively; used with permission.

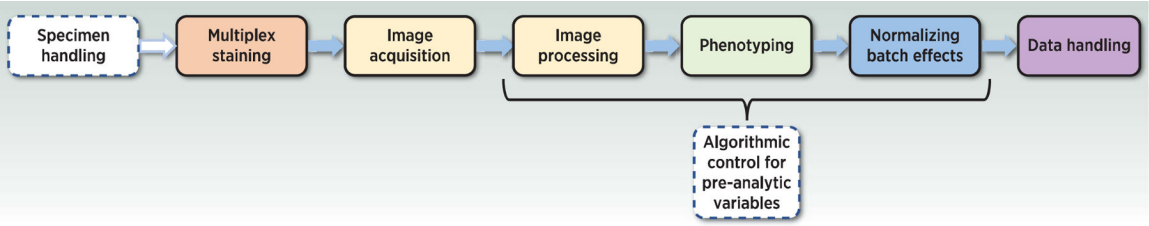

Figure 4. Multiplex immunofluorescence (mIF) AstroPath pipeline for high-quality, whole slide data with single cell resolution.

Each step from staining to batch correction needs to be formalized and standardized to ensure the generation of robust, reproducible mIF data. Batch-to-batch corrections may be facilitated by running a control slide with each batch, or ideally, with a positive control section mounted on the same slide as the patient control specimen. In the future, controlling for pre-analytic variables will further contribute to the reproducibility of results. Specifically, pre-analytic variation may be mitigated by standardizing specimen handling procedures such as establishing maximum cold ischemia time, uniform fixation conditions for tissue, etc. It may also be possible to algorithmically correct for pre-analytic variation during image processing, using sophisticated computational approaches, allowing for specimens that were not handled uniformly to also be analyzed with robust results.

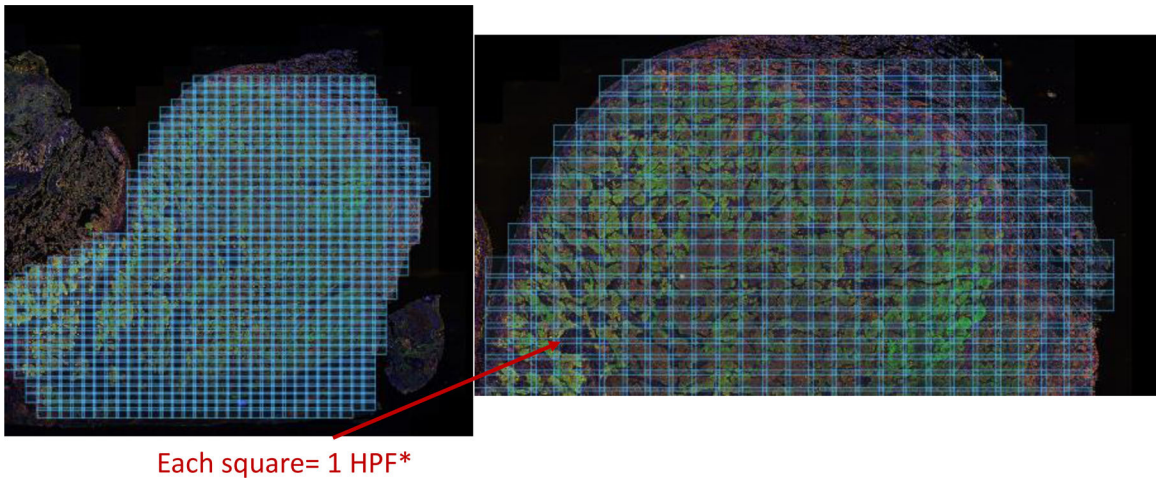

Figure 5. Image tiles of whole slides stained with mIF were acquired using the approach used in SDSS.

Each blue square represents a single HPF image tile. It can take over 1000 HPFs to comprehensively tile the TME on a slide. When stained with ~5–8 mIF markers and imaged in this fashion, each tumor required approximately 300 GB of disk space and benefited from specific data organization and handling considerations learned from astronomy.[27] The resultant mIF maps can be used for biomarker discovery independently, and/or they can also be used to contextualize additional spatial-immune data from other platforms such as spatial transcriptomics or higher-plex mIF of ~40–50 markers, since these other modalities typically image on the order of 1–5% of the tumor surface area shown here in this example.