Leveraging cutting edge tools to convert I-O data into knowledge

Reaching for the stars: combining astronomy and pathology to map cancer biomarkers

Immuno-Oncology Insights 2021; 2(5), 247–256

10.18609/ioi.2021.033

PUBLISHED: 7 SEPTEMBER 2021

INTERVIEW

Janis Taube, Alex Szalay

Janis Taube is the Director of Dermatopathology at Johns Hopkins University SOM, and co-Director of the Tumor Microenviornment Core at the Bloomberg~Kimmel Institute of Immunotherapy and the Mark Foundation Center for Advanced Imaging and Genomics. Her laboratory focuses on characterizing the local, pre- and on-treatment tumor microenvironment in pathology specimens using techniques ranging from routine histology to new multispectral tissue imaging platforms.

Alex Szalay is the Bloomberg Distinguished Professor of Astronomy and Computer Science at the Johns Hopkins University. He is the architect for the Science Archive of the Sloan Digital Sky Survey. His papers span areas from astronomy, spatial statistics, computer science and more recently cancer research. He is a Corresponding Member of the Hungarian Academy of Sciences, and a Fellow of the American Academy of Arts and Sciences. In 2007 he received the Microsoft Jim Gray Award. In 2015 he received the IEEE Sidney Fernbach Award for his contributions to Data Intensive Computing. In 2020 he was awarded the Victor Ambartsumian International Science Prize for his work in physical cosmology. In 2021 he and his team were recognized with the ACM SIGMOD Systems award.

Q Can you tell us a bit about your background and current work?

JT: I trained as a surgical pathologist and am now focused on dermatopathology. This closely ties into the research work that we do, as melanoma was the vanguard tumor-type for immunotherapy Some of the first-in-human trials for immunotherapy were done here at Johns Hopkins, and since melanoma was the tumor type I studied, I had the opportunity to look under the microscope at the melanoma before and after a patient had received immunotherapy, and associate observed features with treatment efficacy. This led to our description of the PD-L1 biomarker, which is now used in numerous FDA-approved tests. It is used as a selection criteria for immunotherapy in more than 40 countries around the world. However, it is imperfect, and one of our goals is to continue to improve upon it. We have now looked at more than 10 different tumor types and used numerous features of the tumor microenvironment to predict response and resistance to these exciting therapies.

AS: I am an astronomer; a Bloomberg Professor of Physics, Astronomy and Computer Science at Johns Hopkins. As a cosmologist, I've worked all my life on the spatial distribution of galaxies, and what can be inferred about the cosmos.

About 25 years ago, I got involved in a project that set out to map the skies – both the stars and the galaxies we can see from the northern hemisphere. I spent a lot of this time building a large spatial database of these galaxy distributions, and then developing spatial statistical techniques to make inferences.

When we met five years ago with Janis, it became clear that the problems for immunotherapy – the interactions of cancer cells and the immune cells – are deep down quite similar to what I have been working on. The galaxies and stars are kind of like the atoms of the structures we see in the sky, and the different types of cells are basically the atoms and structures on the pathology scale.

I have also spent an enormous amount of time on understanding how we can change the way we do astronomy. Astronomy used to be a project for individuals. A single astronomer would point a telescope at a small part of the sky, and study this patch for a large fraction of their lives. Instead, what we generated is the equivalent to something more like the Human Genome Project – we essentially took a census of all the nearest 500 million galaxies on the sky, and then immediately made the data public and available.

This led to an amazing number of discoveries, and over 10,000 refereed papers. Some of the major discoveries were made by computer gamers, or by a middle-school science teacher in the Netherlands. The Galaxy Zoo project, which you may have heard of, also started up in England.

Q How did the AstroPath project come about, and what inspired it?

JT: With all of the technological advances for making scientific observations and improved computing capacity, there are a lot of questions in science that are generating big data. Alex mentioned one such example, which is the genomics and the associated Human Genome Project. There, progress in sequencing led to the generation of large databases and spurred numerous discoveries.

We envision something similar for characterizing the spatial arrangement of cells and the molecules they express in tumors. We used to study factors in the tumor microenvironment one at a time. But it became clear that what we really needed to do was study multiple factors simultaneously, across a broad array of tumor types. New microscopes emerged, which in theory, would allow us to do that in a meaningful way, but we still couldn't manage and query the data that we were generating. We were frustrated as a team, and I was certainly getting pressure for deliverables from those who had helped support the purchase of our new, expensive microscope. Even as I was reporting preliminary results, we were regularly crashing Excel spreadsheets, and it was taking 20 minutes to transfer a single image file from the desktop to the server – and we needed to perform a thousand such analyses to really characterize a single tumor.

It was subsequently suggested to me that Alex Szalay, who was a big data expert here at Hopkins, could help manage all of the fluorescent data and spatial signatures based on his expertise in astronomy.

AS: It was an Easter Sunday, and the President of the university called and told me that there was a new institute being formed around cancer immunotherapy. He saw that there was a good overlap in the techniques that we used to observe the sky, and that I should come to the opening ceremony. The opening speech was given by Joe Biden, and about halfway through his speech he mentioned that he heard that there will be an astronomer applying the techniques used to find patterns on the sky to cancer immunotherapy. I suddenly realized it was me he was talking about! After that we talked with Janis and I met her team – lots of brilliant young kids. It is a very rewarding collaboration.

An important thing about big data is that when you only have, for example, a thousand objects overall, and you find one new object among a thousand, you don't know whether it is significant or what the frequency is. When you have more than 500 million objects, then you can actually understand your data – you can find things that are as rare as one in a million, and you still find hundreds or even thousands of them. This is where we are getting with the cancer data. I expect that by the end of the year, we will probably hit the billion cell mark.

Q Your recent publication in *Science* addresses patient selection for PD-1/PD-L1 therapy in melanoma patients. What are the biggest limitations of current methods, and how does your approach address these?

JT: One of the limitations of the current methods is the fact that most look at one marker at a time, for example, PD-L1 expression; that is standard surgical pathology practice for clinical care and has been for the last 40 years. For research it is possible to visualize multiple markers, and part of what we brought forth with the AstroPath platform is a standardization that should facilitate clinical-grade results, and set the stage for moving multiple markers into the clinical sphere.

Importantly, as a part of our publication we also demonstrated the need for multiple markers. We developed our biomarker using the six factors that we were interested in studying, and we also explored whether we could address patient selection equally well with four factors – and the answer was no.

People have previously stained tumors with multiple markers, but when expression in tissue was assessed, it was scored as positive versus negative. What we were able to do was to look at levels of expression intensity. Not just whether something was 'on' or 'off', but whether it was 'on' at low, medium or high levels. We were able to show the physiologic significance of those differential expression levels and develop a more robust biomarker because of it.

Lastly, pathologists read most biomarkers that are in clinical use visually now – certainly PD-L1 is read using visual light microscopy, and there is variation in interpretation amongst pathologists. What we achieved as a part of the *Science* publication was use the computer to quantitate marker expression. Such digital reads will be really important for standardizing six markers, and we demonstrated feasibility and reproducibility as a part of the study.

AS: Some of the things that we introduced are seemingly simple, but are very important for reproducibility, how we calibrate, and how we intrinsically measure the quality of our data.

One of the ideas also came from sky surveys, where we use overlapping images to build up a mosaic of the sky. The images we took were not just edge-to-edge, but have significant overlap between them. There was about 10–20% of the sky that was observed twice, under independent conditions. By comparing the observations between the data, the fluxes that we measured more than once, the positions of the objects and so on, we gained an extremely good internal determination of the errors in the measurements.

In addition, we tried to introduce a systematic calibration every step of the way in order to trace the errors. This is an extremely complex process. We begin with chemical staining of the slides, we scan and segment them, and then we measure the fluxes. Every step along the way there are small errors, which all accumulate. We believe we managed to significantly reduce these.

The third aspect was that in flow cytometry you get a statistical distribution of the properties of the cells involved, but you don't know where they come from within the tissue. With our approach, for every cell we not only trace its position, but we even trace the precise outlines of the nuclei. We are using a GIS map; (geographic information system). In a sense, our tissue map looks like Google Earth.

We know exactly which cell is physically touching another cell, or which cells are in a close proximity. We can see how these interactions happen, and we can quantify these interactions – not just by manually counting, but by running complex queries over very complicated boundaries.

Q What do you envision next for the platform?

JT: We are increasing our number of melanoma specimens, and we have moved on to lung cancer, pancreatic cancer, and colorectal cancer, among other tumor types. We are asking similar questions: what are factors in the tumor microenvironment that can predict response, and how can we get the right patient on the right therapy? Another big question is how we can look for pan-tumor signatures. What are the commonalities across these different tumor types that provide insight into how the immune system interacts with cancer?

We are also working on adding additional markers to the tumors that are already in the database. We started with six, but it is possible to take an adjacent slice of tumor that's four microns away and build another layer on top of the first slice, looking at different, additional markers. We already have a few additional slices from the cases that were presented in the Science paper. Alex has loaded those into the database, and we are starting to query them. We are really excited about that.

Another major component that we are working on now is looking at on-treatment tumor specimens. We have focused so far on assigning the right therapy to patients as they start treatment, and we are now spending more time looking at what happens in the tumor once patients are on therapy. This gives us some good insights into the mechanism of action of the agent, and what I like to call the necessary 'immuno-architectures' i.e., how the immune cells have to spatially arrange themselves to achieve tumor clearance – they may have a home base of a tertiary lymphoid structure, which then allows sorties to be run from there by executing lymphocytes. Once we know what is needed to achieve tumor clearance, it can inform our understanding of therapeutic resistance and potentially inform future combinatorial treatment regimens.

AS: We built a nice visual interface so that we can see all this spatial data in a flexible way. We can overlay the outlines cells on top of the images and color-code them.

When we first looked at these post-treatment samples, it was striking to see this wave of cells. In the center there are the tumor cells, and surrounding those are the immune cells and a ring of dead cells – cancer cells that have been killed already. It was amazing how the interactions that were taking place became immediately obvious. A good image tells a thousand stories, as they say. But it really took a lot of work to find the right way to present this information visually.

Another thing we are getting increasingly excited about is spatial transcriptomics. There is another deep astronomy analogy here. With our sky maps we took five-color images of the sky first, and the colors of the galaxies already told us a lot about their properties. Then, we took 640 optical fibers and pointed each of them on a different galaxy (within the field of view of the telescope), and then we took a spectrum of it, resulting in more than 4000 resolution elements along the spectrum. This is essentially what spatial genomics does, so that we can look at the region on the tissue and get the mixture of the genomes of all the cells.

When a particular gene is expressed, we don't really know which cell it is coming from. The same thing happens when I place an optical fiber on a galaxy – on the image I can see the fine structure of this galaxy. But when I take a spectrum, I get the sum of all those parts. We are starting to develop techniques to deconstruct how the different morphological components contribute to each part of the spectrum. The same idea is applicable to AstroPath.

This is our vision of how spatial transcriptomics will be applied – that we can eventually associate that this part of the gene expression is most likely coming from the tumor cells, this part is coming from the T cells, and so on. We believe that combining all this information and tying it together will give us a completely new dimension.

JT: In addition to adding spatial transcriptomics into the database, we are adding whole exome sequencing for a lot of these tumors too. The specimens in the database are heavily clinically annotated, and include information on whether a patient responded to therapy as well as long-term survival data, and information on whether they developed immune-related adverse events related to these immunotherapies. The end goal will be to be able to ask multi-modality multiplex questions about factors contributing to the clinical responses that we observe.

AS: In yet another analogy to astronomy, we made all of the data freely available. Over the next year or two we are going to be build an open database, a cancer cell atlas, which is tied to all of the additional information stored at NIH in the TCGA data set.

Q Why are big data and computational biology approaches so significant to immuno-oncology, and what changes or improvements are these approaches currently bringing to the field?

AS: Scientists like to dig into data. But even if the data is big, they still expect to have interactive access and immediate feedback.

We created an environment which makes it very easy for people to share even intermediate results. For example, let us take a simple table that combines two quantities. You compute it, and then you want to share it with your collaborators and see what additional thoughts it inspires. We created an environment where you can do this data markup/fusion very efficiently.

However, although we are getting increasingly into machine learning and artificial intelligence tools, there are still certain steps that require a human pathologist, such as creating the annotations, the tissue outlines, and so on. There is still a human in the loop. This undoubtedly creates some bias. No two people will create exactly the same outlines or segmentations. It would be nice to automate this, but it is very hard work because we have to build up these large datasets in order to train the artificial intelligence tools. We are chipping away at it and we will get there one day – maybe sooner rather than later.

JT: Alex has not only enabled the work, but now that all of this information is in the database, we can explore hypotheses in a rapid-fire way. It is so much fun to be able to do that and get immediate feedback. Of course, as scientists, we are willing to go through the grind of generating good data. But it was truly a grind before. Now, getting to rapidly adapt and ask tons of questions in a very meaningful way makes using these new technologies enjoyable as well as gratifying.

Q What challenges still stand in the way of these approaches fulfilling their potential? What advances would you most like to see?

JT: This isn't directly related to big data, but the multiplexing platforms debuted in a big way in a flurry of proof-of-principle papers. However, I don't think there was appropriate standardization and reproducibility.

My concern is that there will be a continued proliferation of results that aren't necessarily comparable between all of the new and emerging platforms. This gets to what Alex talked about in astronomy: everyone doing their own studies independently, and a lack of cross-institution collaboration or alignment. Our vision is to move multiplex immunofluorescence tests on tumor tissue into clinical care, and before that happens, what we need as a field are agreed-upon, universal standards.

I am not wedded to any one platform, and they all have their own merits, but we need to make sure that the results from them are comparable. We tried to provide the scaffolding for that in our paper by providing suggested universal standards, but ultimately, those will have to be agreed on by the field as a whole.

Once a standardized approach is agreed upon, the next step will be to prove that test results can be reproduced across multiple institutions. We have made the first advances in that direction – we just published a multi-institutional reproducibility study focused on staining tumor tissue with six markers at a time. Now we have started a follow up multiinstitutional study with a focus on slide imaging and analysis. Such studies will need to be conducted for each of the different platforms that moves towards clinical use.

Ultimately, to bring it to clinical care we are going to need regulatory body approvals, and payers who understand the value of multiplex immunofluorescence tests and are willing to reimburse for them. All of that is ahead of us once we generate the next round of reproducibility studies.

AS: As these ideas develop further, people will be using different types of microscopes or flatbed scanners that are taking a slide in a slightly different way, using a different staining mechanism or different markers. We will want to integrate all these data, which means that we have to be much more careful about establishing a common global coordinate system for a slide.

Flatbed scanners can have very different distortions from microscopes, which have a circular objectives, with some pincushion distortions, and so on. These all need to be taken out until we can create a good common reference system.

This multi-modality aspect will be part of our life. The genomics, spatial genomics, and transcriptomics will come in many different forms. Today, there are already multiple instruments which generate data in different formats. We will somehow have to bring them to a common denominator. This field is complex and finding a way to handle these challenges in a coherent and simple fashion will be crucial.

Another issue to consider is funding. In general, using cutting-edge technology for science is not something that is very easily picked up by federal funding agencies or governments. They are playing catch-up, and private foundations are the ones primarily driving it. In astronomy, our project was funded mostly by the Sloan Foundation, and this effort is again funded 90% by private foundations. This is a major change in science.

If the government funding agencies don't wake up and become more agile in following modern, sometimes risky technologies, I am afraid that this may cause some harm when it comes to developing new ideas.

JT: It is really interesting to consider the people who are investing in big data and science as it applies to health. We received support from the Melanoma Research Alliance, Bloomberg Philanthropies and Emerson Collective. Our biggest funding, which really came at a critical time, was through the The Mark Foundation for Cancer Research. All of these entities have connections to entrepreneurial, billionaire investors.

Q Looking to the future, how do you predict that better leveraging big data can impact the field of immuno-oncology?

JT: I think we will be able to identify additional treatment targets in a rational way.

Right now, different treatment regimens are being combined with each other empirically. In the future, we are going to be able to identify additional checkpoints and other immune-active molecules that should be co-targeted, along with PD-1 based on our understanding of their expression patterns within the tumor microenvironment. I think we are also going to gain a lot of insight into the fundamentals of how these drugs work. We have some understanding now – but when you really press, we don't yet fully understand how these agents work. By looking closely at the tumor tissue as it's being cleared, we should be able to gain additional insight into what is happening in the local tumor microenvironment under therapeutic pressure.

AS: To date we have four clinical cohorts in the database. The number of pixels we have to process is already twice the number of pixels that we had to process for the whole sky survey – and that took 25 years and US$250 million! This is moving incredibly fast: within two years we will be scaling-up the number of microscopes and working on automating.

We will be dealing with serious amounts of data. It will not be quite on the level of CERN and the Large Hadron Collider, which has tens of petabytes, but we will probably have multiple petabytes of data to deal with. In contrast, all of the current medical data at Johns Hopkins is of the order of six petabytes – and we just bought a four petabyte storage system for our group.

It is amazing to see how easily the young people we have hired absorb both sides. The people on Janis' side of the group are picking up computer skills, and the kids on my side are learning the biology. The next generation of scientists are becoming more multidisciplinary, and are genuinely excited about these interactions. Our higher education system is suited to create people who drill deeper and deeper down in a single area, but we need people who are reaching across many disciplines at the same time, and who are equally at home in computer science or biology.

Science has gone through multiple paradigms. The first one was just empirically collecting data, with star charts, and Leonardo's codices, and so on. Tycho Brahe then gave all his empirical data to Kepler, and Kepler abstracted them into a set of elegant equations. This led to the emergence of theoretical science. At the same time, science started to fragment into physics, mathematics, chemistry, and biology. Theoretical science culminated with Einstein and the theory of relativity. After that, computers suddenly emerged. People started to use the computers to do calculations and solve some of these analytical theoretical equations, and this brought about the computational era of science. Over the last few years we have seen data-driven discoveries emerging, and this is the fourth paradigm of science.

The empirical era lasted thousands of years, the theoretical one took hundreds of years. The computational era was over in tens of years, and now the data driven era is happening right in front of our eyes. It is just as significant as any of these previous epochs, but there is one new characteristic. Science has been fragmenting into more and more specialized sub-disciplines. With data-driven science, a lot of the techniques are again common, so we see a synergy and convergence. Who would have thought that life sciences and astronomy have so much in common? A new scientific revolution is happening when and where we live, and it's an amazing feeling.

Affiliations

Janis Taube
Director of Dermatopathology, Johns Hopkins University
and
Co-Director of the Tumor Microenviornment Core, Bloomberg~Kimmel Institute of Immunotherapy and Mark Foundation Center for Advanced Imaging and Genomics

Alex Szalay
Professor of Astronomy and Computer Science, Johns Hopkins University

Authorship & Conflict of Interest

Contributions: All named authors take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Acknowledgements: None.

Disclosure and potential conflicts of interest: Dr Taube has a patent issued to multiplex IF assay and associated image processing, stock options for Akoya Biosciences and has received a loan of equipment and provision of reagents from Akoya Biosciences. Dr Szalay has two patents pending and holds stock options for Akoya Biosciences. He holds leadership roles in the National Academies of Science, the Humboldt Association Strategic Advisory Panel on Data, the Heidelberg Institute for Theoretical Studies Scientific Advisory Board and Akoya Biosciences.

Funding declaration: Dr Taube has received research funding from Akoya Biosciences and BMS, and consulting fees from Akoya Biosciences, BMS, Merck, Astra Zeneca and Compugen. Dr Szalay has received support from the NIH, Bloomberg-Kimmel Institute and Emerson Trust. He has received grants from Akoya BioSciences, Melanoma Research Alliance and The Mark Foundation, consulting fees from Akoya BioSciences and payment or honoraria from Genentech.

Article & copyright information

Article source: Invited.

Interview conducted: Jul 14 2021; **Publication date:** Sep 7 2021.

}

BIOINSIGHTS

Connect with us