AGAVE: Automated Genomics Application for Variant Exploration

Hannah M. Gooden, Kenneth V. Bowden, and Brian B. Merritt

ABSTRACT

The Johns Hopkins University Applied Physics Laboratory (APL) is actively developing new capabilities for genomic surveillance of viruses. APL genomicists analyze, process, and visualize viral genomic data for several sponsor organizations that require those data to inform clinical, research, and public policy decisions. Many of the final products from these processes are delivered to sponsors as static reports or slide presentations, but it can be arduous to review or extract pertinent information from these documents. APL genomicists wanted to improve their sponsors' ability to analyze their data and rapidly identify genomic samples or sequences they find important for decision-making. With this goal in mind, a group of APL software engineers developed the Automated Genomics Application for Variant Exploration (AGAVE). AGAVE is an interactive, intuitive web-based tool where researchers can explore and analyze genomic data, draw new connections between data points, and understand the significance behind genomic variants quickly. Researchers can view their sequence data, choose a reference genome with which to compare the data, visualize the 3-D structure of proteins that would be created from particular segments of DNA, and export those visualizations as easily shared image files. AGAVE is still under development and currently supports only influenza genomes, but as it matures and its user base grows, it will expand beyond influenza to include other viruses such as SARS-CoV-2 and even bacterial genomes.

INTRODUCTION

Since the Human Genome Project began in 1990, advances in genomics have fueled explosive growth in the fields of biological and health informatics.[1] The ability to analyze and understand the DNA building blocks of life has led to entirely new fields of scientific research and development, including gene editing, precision medicine, gene therapy, DNA forensics, and more. Additionally, the recent COVID-19 pandemic has brought genomic sequencing technologies into the worldwide public spotlight. After decades of genomic research into the SARS (Severe Acute Respiratory Syndrome) virus, researchers were able to use existing genomic data to

swiftly create revolutionary vaccine technologies for SARS-CoV-2. Alongside the widely publicized contributions of the genomics community, the COVID-19 pandemic triggered a worldwide conversation about the need for additional research into the mitigation of disease-causing pathogens.

APL's Biological Sciences Group has made significant contributions in the field of viral genomics. In collaboration with the Johns Hopkins Center of Excellence for Influenza Research and Surveillance (JH-CEIRS), we conducted genomic surveillance of the influenza virus over several years and applied that expertise to surveillance of SARS-CoV-2 in the US National Capital Region at the beginning of the outbreak in 2020.[2] APL has added significant amounts of data to the body of knowledge that assists with flu vaccine development and clinical and public policy decisions. APL has in-house sequencing capabilities and has developed a robust catalog of pipelines to analyze genomic data and generate information for use by clinicians, researchers, and public policy decision-makers. APL-developed pipelines for DNA processing, including an open-source desktop application known as Basestack, ease distribution and installation requirements for health organizations in need of in-field research applications. Basestack has enabled genomic sequencing and analysis in numerous countries, including Bangladesh, Pakistan, Sierra Leone, and Chile. In 2021, Fast Company recognized APL as one of the top 20 workplaces for innovators, commending the Lab for focusing "its prodigious resources on refining real-time DNA sequencing for new viruses."[3]

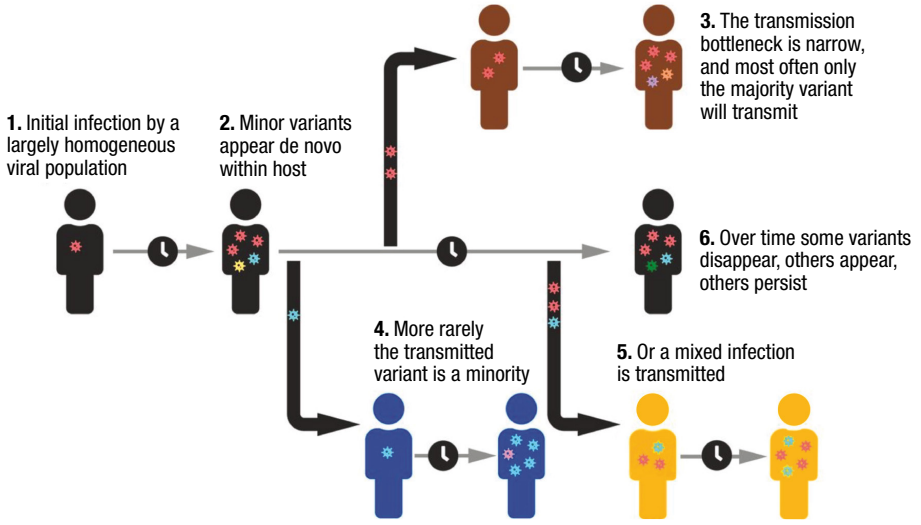


Figure 1. Flowchart showing the method by which minor variants develop among populations. The genetic diversity of an initial infection in a single host is typically very low, but as the virus reproduces, minor variants begin to appear. The host will transmit both the major and minor variants, causing viable minor variants to slowly grow in prevalence within a population. (Figure from Lythgoe et al.,[5] FontAwesome, licensed under CC BY 4.0; https://creativecommons.org/licenses/by/4.0/.)

PROBLEM

Through developing these capabilities, we recognized that visualization of genomic data is a space with tremendous opportunity for innovation. We began asking bigger questions: How could we develop creative methods of delivering data to sponsors and clinicians in ways that would be readily digestible and easy to replicate? Could we streamline the process of exporting useful insights as graphics that could be easily integrated into presentations and publications? Standard deliverables involved arduous pipeline development and support to end users. Additionally, we provided data to sponsors in high-level reports, often in the form of static documents or slide presentations. We had initial success in leveraging Basestack to deliver powerful narratives by allowing researchers to explore data and results through fully interactive and reproducible graphics. We wanted to build on this success by continuing to use web-based technologies to change the way our sponsors interact with data. From these conversations came AGAVE: the Automated Genomics Application for Variant Exploration.

AGAVE is an avenue for exploration of minor variants within DNA or RNA viral genomic sequences. In genomics, minor variants "are defined as the variants with frequencies lower than 10% in a cell population."[4] Minor variants are clinically significant because they allow researchers to understand the genomic diversity of the whole population of organisms within a single sample (Figure 1).

For clinicians, minor variant analysis is an effective and necessary step in preventing infection breakthroughs within vaccinated populations and improving vaccine development research. To detect minor variants in sequenced DNA or RNA, clinicians follow a process of workflows using bioinformatics tools, and then computational biologists manually inspect the results. The specific location of minor variants on a gene is significant because genes encode different proteins that determine factors such as virulence and transmissibility. Minor variants occurring on certain genes may lead to a dramatic shift in vaccine efficacy and even evasion

of immune responses in the host, so a minor variant occurring in a gene that encodes a clinically significant protein may be of utmost importance for a researcher to analyze. An example is the hemagglutinin gene, which encodes for the protein that covers the surface of a flu virus and allows the virus to enter the host cells.[6] For this reason, hemagglutinin variants are primarily considered when choosing flu strains for vaccine production. The ability to interactively explore segments and determine the variants on key proteins like hemagglutinin provides a crucial understanding of how these variants would affect the physical structure of the virus and ultimately its ability to infect the host and reproduce. The resulting insights of this type of analysis will enable researchers to identify new strains of a virus that could be included in yearly vaccines.

This analysis is critical, and the format in which the resulting data are delivered is important as well. Many bioinformatics tools export raw data in custom file formats, leaving it to the researcher to format results in useful and visually intelligible ways. Researchers spend a lot of time manually creating static and heavily curated visualizations for specific results, drastically increasing the time required to deliver useful information to a wider audience. Thus, an interactive tool that enables users without a bioinformatics background to easily access this information and includes functionality for exporting data would be valuable to the field.

We approached both these goals with AGAVE, a tool created for effortless exploration of genomic data, allowing any clinician, researcher, or policy decision-maker to assess viral changes in the regions they most care about. In addition, we included features within that tool to easily export data as graphics for reporting or inclusion in scholarly work.

AGAVE APPLICATION

AGAVE is an interactive web application that provides several ways for a researcher to explore their viral genomic sequence data (Figure 2). It consists of three primary components: a heatmap that shows positions with a high proportion of mutations at a glance, a molecule viewer that renders the selected protein in three dimensions, and a stacked bar chart that displays more detailed visualizations of genomic diversity at specific positions. All components are linked dynamically, assisting users in conceptualizing the impact of each positional variant on the results. To prepare genomic data for visualization by AGAVE, DNA samples should be extracted from biological material(s), sequenced, and analyzed using techniques that include variant determination. These preparatory steps can be performed using sequencing equipment located on-site at APL and analysis pipelines developed by APL genomicists. The prepared data can then be loaded into AGAVE for visualization via its three primary components.

The heatmap (Figure 3) is the largest and most important component of AGAVE because it creates a high-level view of the entire data set. This view helps

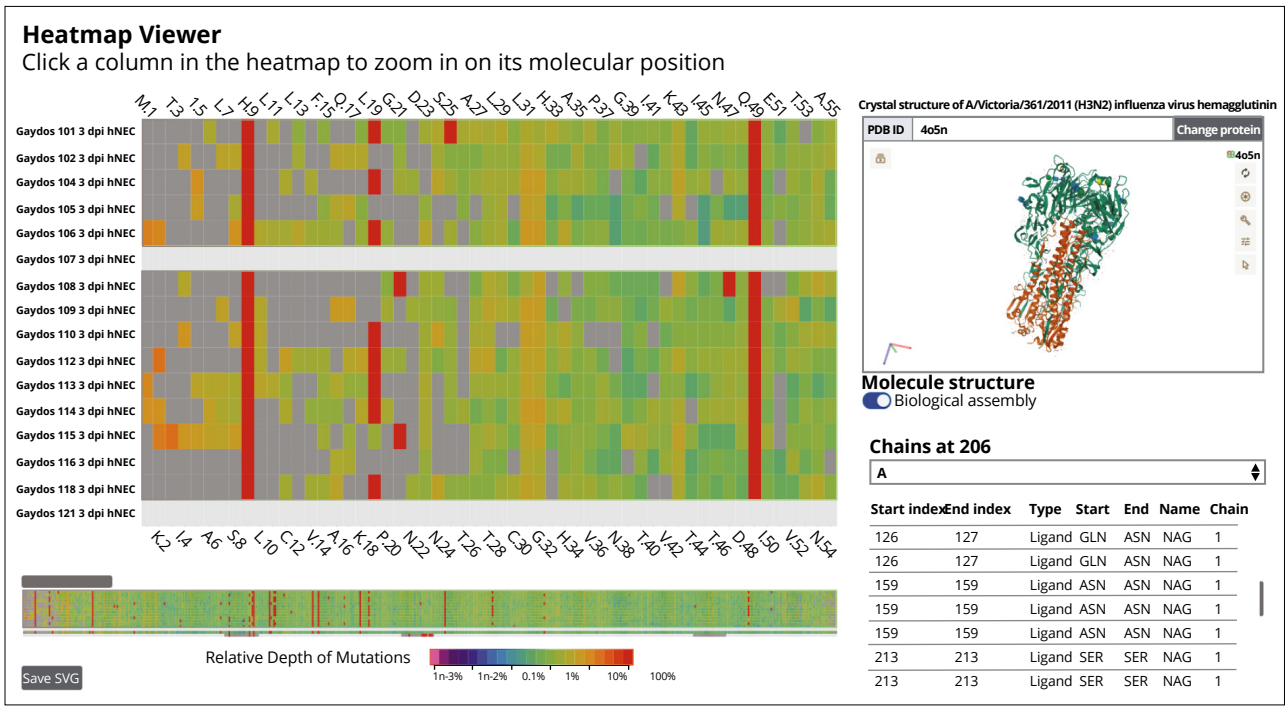


Figure 2. An overall view of the AGAVE application. This view shows the heatmap and molecule viewer next to each other on a single screen.

the user rapidly identify genomic regions in need of further investigation. Researchers can select DNA samples to explore in the Settings panel (Figure 4), and those samples will be added to the heatmap's *y* axis. Each position along the samples' genomes is plotted along the *x* axis, and the box where each sample name and position number intersect is colored according to the number of mutations observed. Figure 3 depicts how all positions are colored according to the mutation frequency, from proportionally less to more—violet to red, respectively. This color scheme allows researchers to quickly scroll through their data and observe where intense pockets of mutations are occurring. They can also hover over positions individually to see what minor variants are occurring.

The heatmap is also highly customizable; for example, researchers can change whether variants are calculated based on a known reference genome or on the individual assembled consensus of each sample, which is derived from several bioinformatic tool kits that best assemble a genome based on the many fragments of DNA created during the sequencing process. In addition, AGAVE includes flu-specific settings, so users can select which flu subtype their data should be compared with, as well as which segment of the flu genome they would like to view. In future work, we look to target more organisms of interest, including bacteria or other viruses.

Next to the heatmap is a 3-D representation of the assembled protein (Figure 5). This rendering is an interactive 3-D model of the protein produced from the genome segment the user is viewing in the heatmap. For example, if the user is viewing data from the gene that encodes the neuraminidase protein, the molecule viewer will automatically bring up a 3-D model of neuraminidase. The protein is also linked to the heatmap, so that if a user

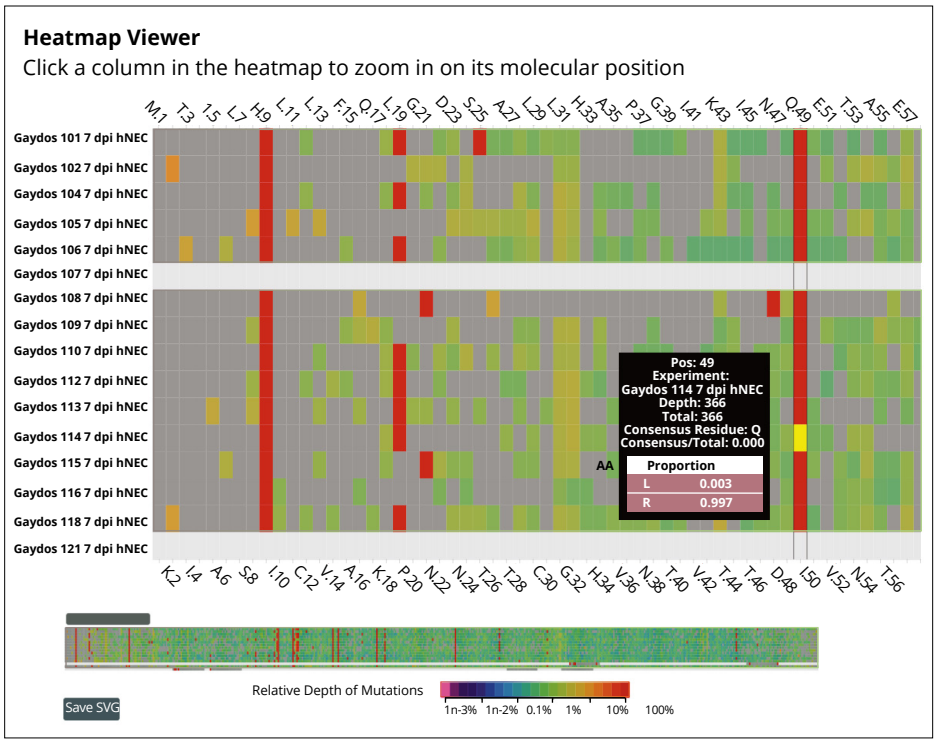


Figure 3. The heatmap viewer. This component plots viral genomic samples on the *y* axis and positions on the *x* axis and shows relative depth of mutations as color. Along the bottom, the entire length of the genome is shown in a summarized form. This figure shows the heatmap focused on a single position, with a tooltip giving more information about the data at that position. This tooltip is automatically generated when a user hovers over a heatmap square with their mouse.

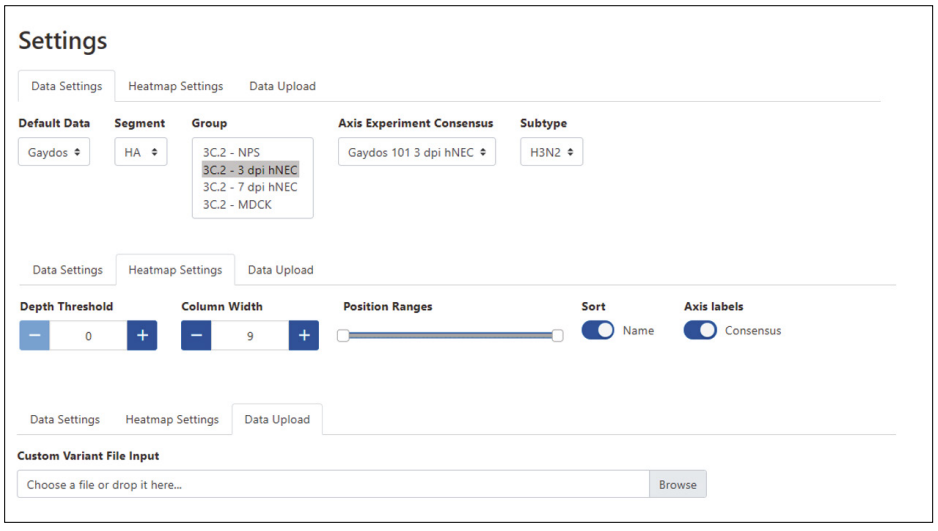
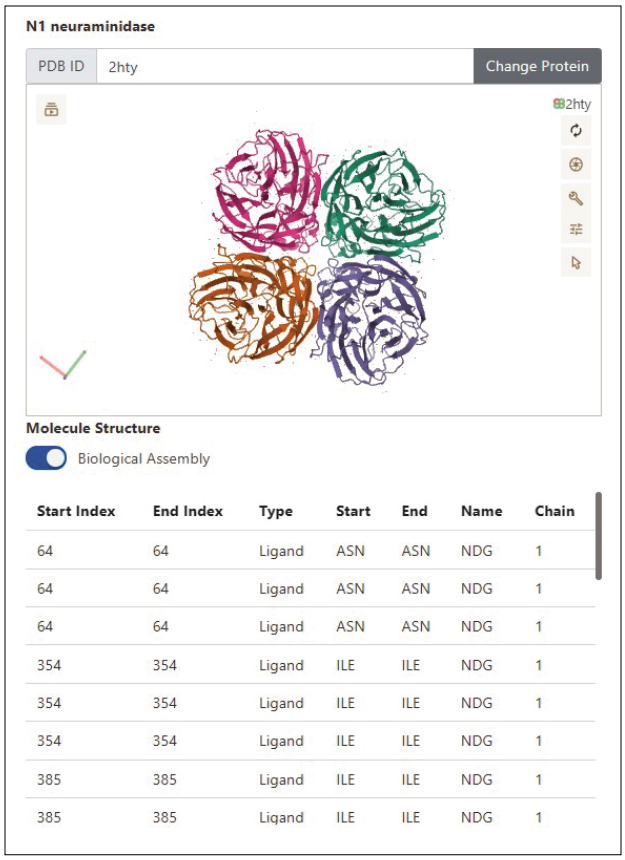


Figure 4. AGAVE Settings panel. AGAVE offers a multitude of settings for creating a custom heatmap view. The settings are divided into three tabs: Data Settings, Heatmap Settings, and Data Upload. The contents of each tab are shown.



Start Index	End Index	Type	Start	End	Name	Chain
64	64	Ligand	ASN	ASN	NDG	1
64	64	Ligand	ASN	ASN	NDG	1
64	64	Ligand	ASN	ASN	NDG	1
354	354	Ligand	ILE	ILE	NDG	1
354	354	Ligand	ILE	ILE	NDG	1
354	354	Ligand	ILE	ILE	NDG	1
385	385	Ligand	ILE	ILE	NDG	1
385	385	Ligand	ILE	ILE	NDG	1

Figure 5. The molecule viewer. This viewer renders a 3-D model of the protein associated with the genome segment the user is currently viewing in the heatmap. A user can rotate and zoom in on the protein to explore its structure. Hovering and clicking on amino acid residues in the 3-D rendering will reveal more information about the individual amino acid residues at that location. We used PDBe Molstar to create this molecule viewer.[7]

clicks on a position in the heatmap, the molecule viewer will zoom in on the corresponding amino acid residue within the protein. As a user scrolls through the heatmap, identifying mutations of interest, they will easily be able to view the corresponding positions of these mutations on the physical structure of the protein.

Below the heatmap and molecule viewer, the variation at single positions across all samples will be visualized in a stacked bar chart (Figure 6). As they can with the molecule viewer, users can click a position in the heatmap to view localized metrics via this chart.

Finally, we ensured that AGAVE produced exportable graphics. Each component has functionality for saving the current view as an image file. Because each component is configurable, researchers can experiment with settings until they have found a visualization that meets a specific need and save the visualization to their file system for distribution to other teams or for placing in reports and manuscripts.

METHODS

We drew inspiration for AGAVE from an application designed years ago by other APL researchers within the Research and Exploratory Development Department's Biological Sciences Group. They created a web application similar to AGAVE that could explore minor variants in data that were analyzed in-house.

We designed a new application built on the same data visualization ideas as the original code base, reusing code where we could but focusing on creating something that would be easy for researchers and clinicians to use. The new design included enhanced interactivity and updated web technologies. Since AGAVE was funded by the Biomedical Advanced Research and Development Authority (BARDA) for a project that primarily sought to analyze influenza genomes, the structure of an influenza genome inspired our web design. However, our design is not exclusively tied to influenza, so it remains flexible and reusable.

So far, we have completed the AGAVE user interface. We wrote the front end using TypeScript[8] and the framework Vue.js,[9] and we embedded visualizations that use D3.js[10] and PDBe Molstar.[7] These languages and tools were already familiar to the developers on the project, allowing for rapid prototyping and very quick development cycles. Currently, the data AGAVE consumes have been formatted and loaded manually. We are looking to build out a back end for AGAVE using Express[11] and incorporate data analysis pipelines, which can be deployed remotely and linked to other APL-built tools, including Basestack, mentioned earlier in this article.

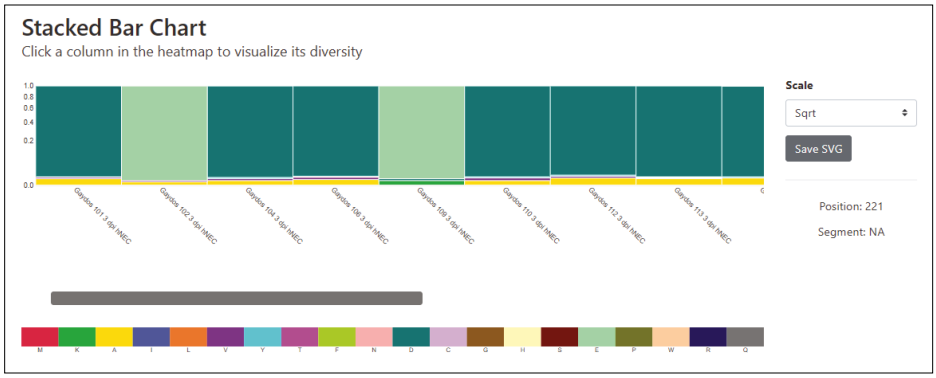


Figure 6. The stacked bar chart. This view shows the incidence of mutations along a single position across samples currently shown in the heatmap.

FUTURE DEVELOPMENT

Although AGAVE is still in its proof-of-concept stage, it has already demonstrated its value. APL researchers in the Biological Sciences Group have been using AGAVE to explore flu sequencing data for BARDA. By loading sequenced influenza genomes and in-house analysis into AGAVE, they were able to find examples of minor variants that they would have otherwise missed, and they then used these minor variants to build a story around which to frame their data analysis.

The next steps for AGAVE are twofold. First, we hope to broadly generalize its function so that it can be used to analyze variants in any genome type. Currently, AGAVE has been used only with influenza, but it has potential to make an impact on analysis of other organisms, including SARS-CoV-2. Second, we will look to integrate AGAVE with the other capabilities APL has developed over the past few years. We believe AGAVE and Basestack could someday be bundled together for users, where Basestack analyzes genomic data that users input and AGAVE interactively visualizes the result.

AGAVE shows great promise as a new tool for genomic surveillance. We hope to expand AGAVE for broad use within APL's Biological Sciences Group as a user-friendly tool to explore genomic data and add to the body of knowledge within the genomic surveillance ecosystem.

REFERENCES

[1]K. A. McCormick and K. A. Calzone, "The impact of genomics on health outcomes, quality, and safety," *Nursing Manage.*, vol. 47, no. 4, pp. 23–26, 2016, https://doi.org/10.1097/01.numa.0000481844.50047.ee.

[2]P. M. Thielen, S. Wohl, T. Mehoke, S. Ramakrishnan, M. Kirsche, et al., "Genomic diversity of SARS-CoV-2 during early introduction into the Baltimore–Washington metropolitan area," *JCI Insight*, vol. 6, no. 6, art. e144350, 2021, https://doi.org/10.1172/jci.insight.144350.

[3]Fast Company, "Best workplaces for innovators 2021," Aug. 4, 2021, https://www.fastcompany.com/best-workplaces-for-innovators/2021.

[4]Z. Feng, J. C. Clemente, B. Wong, and E. E. Schadt, "Detecting and phasing minor single-nucleotide variants from long-read sequencing data," *Nature Commun.*, vol. 12, no. 1, art. 3032, 2021, https://doi.org/10.1038/s41467-021-23289-4.

[5]K. A. Lythgoe, M. Hall, L. Ferretti, M. de Cesare, G. MacIntyre-Cockett, et al., "SARS-COV-2 within-host diversity and transmission," *Science*, vol. 372, no. 6539, art. eabg0821, 2021, https://doi.org/10.1126/science.abg0821.

[6]J. J. Skehel and D. C. Wiley, "Receptor binding and membrane fusion in virus entry: The influenza hemagglutinin," *Annu. Rev. Biochem.*, vol. 69, no. 1, pp. 531–569, 2000, https://doi.org/10.1146/annurev.biochem.69.1.531.

[7]"PDBeurope/pdbe-molstar." GitHub. https://github.com/PDBeurope/pdbe-molstar (accessed Nov. 3, 2021).

[8]"TypeScript." Microsoft. https://www.typescriptlang.org/ (accessed Nov. 3, 2021).

[9]E. You, "Vue.js," https://vuejs.org/ (accessed Nov. 3, 2021).

[10]M. Bostock, "Data-driven documents," D3.js, https://d3js.org/ (accessed Nov. 3, 2021).

[11]"Express.js." OpenJs Foundation. https://expressjs.com/ (accessed Dec. 2, 2021).

Hannah M. Gooden, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Hannah M. Gooden is a software engineer in APL's Research and Exploratory Development Department. She holds a bachelor of science in computer science from Texas A&M University, with minors in astrophysics and cybersecurity. Hannah is interested in projects at the intersection of software engineering, data visualization, and scientific research. As part of the Discovery Program, she has sought out projects that unite these three fields and provide her with new professional experiences. Her projects have included graphical user interface development for a quantum computing team, ground software engineering for several APL space missions, and cyber situational awareness for several military sponsors. Her email address is hannah.gooden@jhuapl.edu.

Kenneth V. Bowden, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Kenneth V. Bowden is a molecular and genomics scientist in APL's Research and Exploratory Development Department. He holds a bachelor of science in biology and a bachelor of arts in economics from the University of Maryland, Baltimore County and a master of science in biotechnology from Johns Hopkins University. Ken is responsible for end-to-end processing of genomic data, from physical sample preparation and sequencing to bioinformatics analysis. He works on a variety of sponsored projects involving a broad range of organisms, including viruses, bacteria, and plants. His email is kenneth.bowden@jhuapl.edu.

Brian B. Merritt, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Brian B. Merritt is a bioinformatician and software developer in APL's Research and Exploratory Development Department. He holds a bachelor of science in biochemistry from the University of Georgia and a master of science in bioinformatics from the Georgia Institute of Technology. Brian primarily develops a suite of computational biology and web-based tools aimed at improving the field of genomics and sequencing-based research and applications. These include desktop applications, visualization techniques, and hardware productions and developments. His email is brian.merritt@jhuapl.edu.