Next-Generation Sequencing Data Analysis, Management, and Storage in the Cloud

BaseSpace™ Sequence Hub increases an Oxford laboratory's productivity and cost-effectiveness in meeting the needs of its clinical research teams.

Introduction

Today's next-generation sequencing (NGS) systems generate enormous amounts of data, causing researchers to wonder how to manage, analyze, and store all that information efficiently. The cloud provides a solution, offering unlimited data storage, real-time sequencing run monitoring, and access to powerful data analysis tools. For Helene Dreau, MSc, Principal Clinical Scientist at the Haemato-Molecular Diagnostics Laboratory at the Oxford Molecular Diagnostic Centre (OMDC), it also reduced the need for creating a bioinformatics service.

Ms. Dreau and her five-person team are responsible for supporting the genomic efforts of large clinical and research groups at the University of Oxford and Oxford University Hospitals National Health Service (NHS) Foundation Trust. With the Illumina MiSeq™, HiSeq™ 2500, and HiSeq 4000 Systems, she turned to the BaseSpace Informatics Suite for data analysis. As the volume of data generated by her group's sequencing systems grew, Ms. Dreau transitioned to BaseSpace Sequence Hub for storage, collaboration, and data management in the Amazon Web Services Cloud.

iCommunity spoke with Ms. Dreau about her decision to move NGS data analysis to the cloud and the benefits it has provided her laboratory.

Q: What clinical research teams at Oxford does your laboratory support?
Helene Dreau (HD): We perform testing services associated with hematological diseases (hemoglobinopathies, hemophilia, leukemia, lymphoma, etc.) at Oxford Hospital. We offer specialized testing, including DNA and RNA sequencing, and flow cytometry. We are also part of the Thames Valley Cancer Network and responsible for teaching specialty registrars (physicians) before they move to regional hospitals where they complete their training.

We support clinical research studies at Oxford Hospital. We also collaborate with the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre, which is focused on driving innovation in the prevention, diagnosis, and treatment of disease, and translating advances in biomedical research into clinical practice. We develop and validate new technologies to support these efforts. Over the last decade, funding from these entities has enabled us to acquire the latest tools and technologies to become a genomic center.

Q: When did your lab begin using NGS and what types of sequencing applications do you perform?
HD: We acquired a MiSeq System in 2013, and later added HiSeq 2500 and HiSeq 4000 Systems. We use targeted sequencing panels, such as TruSeq™ Custom Amplicon*, and are developing a translocation panel. We also perform RNA-Seq.

Q: What is your assessment of the MiSeq and HiSeq systems in your laboratory?
HD: Our Illumina NGS systems are performing well and I'm happy with them. I've started using the Illumina Concierge service for the design of targeted panels. I have obtained improved sequencing efficiency and more reliable results using the targeted panels service on the MiSeq System. Efficiency is critical for me, because the MiSeq System is the workhorse of our lab. The process of setting up assays is simple, and it's easy to clean and maintain the instruments.

Q: Why haven't you added someone with bioinformatics expertise to your lab team?
HD: We don't have the budget to hire bioinformatics staff. Even if we did, it's difficult to attract bioinformaticians to an NHS laboratory. If they are good, they want to be in academia where they can publish papers or in industry where they can earn higher salaries. Experienced bioinformaticians aren't interested in a position where they are providing a routine service, running a data analysis pipeline, and assessing and disseminating the results.

Helene Dreau, MSc, is a Principal Clinical Scientist at the Haemato-Molecular Diagnostics Laboratory at the Oxford Molecular Diagnostic Centre (OMDC).

*TruSeq Custom Amplicon has been discontinued. The recommended replacement is the AmpliSeq™ for Illumina Custom DNA Panel.

Q: When you first acquired the MiSeq System, how did you analyze and manage data?
HD: We used BaseSpace Software* and the MiSeq Reporter software on the MiSeq System for data analysis.

Q: How did your data analysis and management needs change after the addition of the HiSeq 2500 and 4000 Systems?
HD: The amount of NGS data increased dramatically after we added the HiSeq 2500 and HiSeq 4000 Systems. We've also seen a 20% increase in targeted panel sequencing requests for clinical research studies and more interest in genomic testing by our other Oxford partners.

> "BaseSpace Sequence Hub enables us to analyze, store, and disseminate data without the need of a bioinformatics staff or a server. It has also supported our increased data analysis workload."

Q: What options did you consider to meet the increased demand for NGS data analysis, management, and storage?
HD: We built a business case for having a server onsite, but found that it was beyond our budget and would require a change in infrastructure to support. In addition, the ongoing maintenance of the servers was cost prohibitive. We decided to begin performing analysis using BaseSpace Apps in BaseSpace Sequence Hub in 2015. In August 2016, we transitioned to BaseSpace Sequence Hub with an Enterprise domain that offers expandable storage (> 1 TB) and 24 hours of bioinformatics professional services support. BaseSpace Sequence Hub enables us to analyze, store, and disseminate data without the need of a bioinformatics staff or a server. It has also supported our increased data analysis workload.

Q: How do you manage access to your research NGS data?
HD: Using BaseSpace Sequence Hub, we have one platform for data analysis, storage, and distribution, and that makes sharing easier and more cost effective. I can create several work groups for different research projects and provide access to specific users. Researchers like the fact that they can view their data remotely.

BaseSpace Sequence Hub also enables me to control data access, and maintain our research project data separately. It's important that we maintain privacy of our research work.

**BaseSpace Software was a precursor to BaseSpace Suite.

Q: How does your team perform data analysis in the cloud?
HD: We run the panels, select and run the appropriate BaseSpace Apps, and then conduct a technical review of the data generated. We're all molecular biologists, not trained bioinformaticians. We're pleased at how easy it is to set up BaseSpace Apps and perform data analysis.

Q: Are there other benefits of managing projects in BaseSpace Sequence Hub?
HD: BaseSpace Sequence Hub provides us with remote access to our NGS data, which is useful when we need results back quickly or when we're away from the lab. When a run finishes on a Saturday afternoon, we can remotely start the pipeline and analyze the data in the cloud using the Integrative Genomics Viewer (IGV) and VariantStudio software. BaseSpace Sequence Hub enables us to keep projects moving and be responsive, even over weekends or when we're at offsite meetings.

With remote access, we can also look at run trends easily. If we see something odd, we can use the Sequencing Analysis Viewer (SAV) software and share data with Illumina Tech Support in the cloud. It enables the Tech Support team to identify the issue quickly and send an engineer if necessary. It's been useful for us in instrument management.

> "BaseSpace Sequence Hub provides us with remote access to our NGS data...it enables us to keep projects moving and be responsive, even over weekends or when we're at offsite meetings."

Q: Which BaseSpace Apps are you using?
HD: The number of apps offered on BaseSpace Sequence Hub is good and they cover many aspects of our work. We use the TSCA app for TruSeq Custom Amplicon panels. For development and validation work, we use FASTQC to check NGS data, and the FASTQ Toolkit to manipulate FASTQ files and perform adapter and quality trimming, length filtering, and down sampling. There are some applications that we can pull down in the pipeline for a quick look at the data, and then use a fancy pipeline later for analysis. For whole-genome sequencing (WGS), we use the Illumina Whole Genome Sequencing App. For tumor-normal subtraction, we use the Tumor Normal App, which uses Strelka Somatic Variant Caller to call somatic small variants, structural variants, and copy number alterations (CNA) in tumor-matched samples.

Q: Does BaseSpace Sequence Hub support International Organization for Standardization (ISO) certification?
HD: As an NHS user, I must be ISO accredited and everything I do needs to be verified or validated. To validate my processes, I have to test all the parameters and make sure that the system

doesn't break. If I'm using a system that is provided by a manufacturer, such as BaseSpace Sequence Hub, the responsibility of validating that pipeline is on the manufacturer. All I have to do is verify it works.

The United Kingdom Accreditation Service (UKAS) also requires that the suppliers I use are accredited against all relevant standards. The ISO 27001 certification for BaseSpace Sequence Hub removes the need for me to demonstrate that QA has been implemented for this step in our workflow.

Q: Would you be able to provide the services you offer without BaseSpace Sequence Hub?
HD: Without BaseSpace Sequence Hub, it would have taken us longer and it would have cost more to get to this level of data output and operational efficiency. We would have needed to recruit a bioinformatician that was interested in setting up processes and systems to store, manage, and share NGS data. With BaseSpace Sequence Hub, we were able to streamline our data analysis and increase our productivity cost effectively, while providing researchers and clinicians with a secure way to access their data.

> "Without BaseSpace Sequence Hub, it would have taken us longer and it would have cost more to get to this level of data output and operational efficiency."

Q: What should laboratory managers consider before moving their NGS data into the cloud?
HD: There are many things to consider if someone's thinking of using the cloud for NGS data analysis, storage, and distribution. They need to determine how much data they will generate, what types of analyses they will be performing, and how long they will need to retain the data. The expense of working in the cloud involves more than just the cost of the license. It also includes the cost of storage and computational time. It surprises people how much data are generated by an NGS run. They don't realize that they will receive FASTQ, BAM, and VCF files, which take up storage space in the cloud, increasing costs. They need to calculate whether the benefits that they receive exceed the cost of creating the bioinformatics framework themselves. In our case, using BaseSpace Sequence Hub is a cost-effective way of analyzing, storing, managing, and sharing the NGS data we are generating. The cost of buying a server, even just for our clinical work, and hiring a bioinformatician is cost prohibitive for us in the current budget climate.

Information governance is also an issue. For targeted panels, the data we obtain is anonymized to maintain patient confidentiality. For the WGS service, concerns remain regarding maintaining confidentiality. Currently, we are overcoming this issue by obtaining participant consents, but moving forward, this will be a challenge.

Q: How do you see your lab growing in the future?
HD: We are creating more targeted panels and want to establish our own private, clinical WGS practice. We are developing cell-free DNA work for prenatal applications and early detection of tumor and minimal residual disease (MRD). We also plan to integrate our WGS data with RNA-Seq.

Q: Can you perform the data analysis for those applications in BaseSpace Sequence Hub?
HD: We have a few analysis pipelines that we've developed that are in BaseSpace Sequence Hub. Because our Enterprise account gives us 24 hours of bioinformatics professional services support, we'll be working with Illumina to develop several new BaseSpace Apps.

Learn more about the Illumina systems and products mentioned in this article:

BaseSpace Sequence Hub, www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub.html

BaseSpace Apps, www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps.html

MiSeq System, www.illumina.com/systems/sequencing-platforms/miseq.html

HiSeq Systems, www.illumina.com/systems/sequencing-platforms/hiseq-2500.html

AmpliSeq for Illumina Custom DNA Panel (replaces TruSeq Custom Amplicon), www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/ampliseq-custom-dna-panel.html

Illumina, Inc. • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com