Heads in the Cloud: Harnessing the Healing Power of Bioinformatics

Drs. Elaine Gee and Frederick Strathmann are leveraging the power of the cloud to bring next-generation sequencing (NGS) to a new frontier. This involves deploying scalable pipelines built on best practice strategies and state-of-the-art technology to perform bioinformatics in the cloud.

Since the first draft of the human genome was published in the scientific journal Nature in 2001, current DNA sequencing technology (NGS) needs a complex pipeline transforming raw sequencing data into clinically actionable results.

ARUP Bioinformatics spans five areas of clinical focus—molecular genetics, immunology, hematopathology, anatomic pathology, and infectious diseases. Application of NGS technology to these areas has resulted in massively complex data needs with issues in high-throughput data storage, workflow management, and standardization of computation.

Drs. Gee (**EG**), director of Bioinformatics, and Strathmann (**FS**), who is part of Toxicology and Mass Spectrometry oversight, discuss clinical bioinformatics. Their 14-member team includes Brett Kennedy, PhD, (associate director, Bioinformatics) and Mark Monroe (lead data engineer).

Expert Edge

Frederick Strathmann, PhD
Acting Scientific Director, BioComputing

Elaine Gee, PhD
Director, Bioinformatics

Q: Why do we need clinical NGS testing?

EG: It has been 15 years since the draft release of the human genome, and scientists do not have a full grasp on the genetic secrets hidden within the 3 billion base pairs of DNA sequence. However, great strides have been made to extract medical utility from a subset of genes. Trials like NCI-MATCH help expand the national "cancer knowledge network" that enables labs to provide state-of-the-art laboratory-developed tests that interrogate medically relevant genes.

FS: NGS is redefining approaches to medicine and is at the forefront of the precision medicine initiative. The amount of data generated per patient is unparalleled, complicating research, test development, and clinical interpretation. With significant efforts being focused on sequencing efficiency and bioinformatics, we need highly trained sequence analysts and pathologist to provide actionable reports for continued scalable success.

Q: What role does bioinformatics play in clinical NGS testing?

EG: Conversion of NGS data into interpretable data requires a complex analytical process. NGS is a massively parallel technique that generates short sequence "reads" from a library of sheared (fragmented) genomic DNA. The bioinformatician reconstructs the genetic information by piecing together these short sequences using various analytical techniques and algorithms. This in essence is the first step of the NGS bioinformatics process. For a majority of our techniques, bioinformaticians build pipelines to convert data derived from genes into human interpretable data through three major steps: sequence alignment and polishing, variant calling, and variant annotation.

FS: Bioinformatics is at the heart of the successful application of NGS technology in medicine. Without the bioinformatics aspect, the sequencing sits in limbo. It is an integrated process that involves numerous technologies and highly skilled people to provide actionable information to physicians.

Q: What does the future entail for bioinformatics at ARUP and possibly elsewhere?

EG: As the cost per genome continues to drop, the sheer volume and complexity of sequencing data generated will explode over time. For example, 60 exome datasets alone will generate approximately 1 TB of input data.

Bioinformatics is turning to the cloud to relieve the bottleneck of limited compute capacity and data storage. The underlying infrastructure we are building to transition to cloud computing is developed in-house using state-of-the art components, including the Snakemake workflow management system, the Docker platform, SaltStack, RabbitMQ messaging, and the Celery-distributed task queue to create a coordinated system that is both standardized and modular for customizability.

FS: The need to scale all aspects of NGS testing is looming. There are currently no generalizable best-in-practice guidelines, but partnerships are forming that will allow a better vision of how these complex data problems will be solved as a community. The infrastructure and modular approach being implemented at ARUP represent a move towards robust bioinformatics solutions to support high-quality, high-volume NGS testing.