WHITE PAPER

# Advances in Sequencing Drive Plant and Animal Genomics to the Cloud

## Bioinformatics

Drought-tolerant plants? Bioenergy crops? Precision medicine? Long-read sequencing and cloud computing are key enabling technologies.

**Ketan Paranjape**
General Manager,
Life Sciences and Analytics,
Intel Corporation

**Steve Hebert**
CEO, Nimbix

**Jason (Chen-Shan) Chin**
Senior Director, Bioinformatics,
Pacific Biosciences

### Introduction

From population growth to climate change, genomic analysis holds a key to understanding and managing a wide range of crucial challenges. Yet despite more than a decade of next-generation sequencing (NGS), addressing these challenges—and taking advantage of transformative opportunities such as precision medicine and bioenergy crops—will require continuing progress in both sequencing technology and data-intensive computational processing.

Long-read sequencing represents a significant advance in sequencing technology, enabling more complete and accurate assemblies of larger, more complex genomes. Long-read is highly data-intensive, requiring significant computing resources within the sequencer itself, as well as in the post-sequencer pipeline to assemble the genome; to analyze, compare, and explore the results; and ultimately to achieve and share novel insights. The output of next-generation and long-read sequencing systems is outpacing even the steady performance increases predicted by Intel co-founder Gordon Moore, making high-performance computing (HPC) resources a critical gating factor for continued success.

Cloud computing has established itself as a cost-effective way to deliver the necessary computing resources. Internal private clouds can add flexibility and reduce costs for large labs that have extensive local infrastructure. External cloud service providers offer these flexibility and cost benefits for smaller labs that are outgrowing their workstation-level solutions and for labs of any size that need more capacity to meet the rising demands.

Intel is actively working to accelerate life science (LS) breakthroughs by understanding industry requirements, designing technologies to meet them, and working collaboratively to create end-to-end solutions. This paper describes long-read sequencing solutions from Pacific Biosciences of California, Inc., and cloud computing from Nimbix based on the Intel® Xeon® processor E5 family. Focusing on plants and livestock, it highlights the use of these technologies at the United States Department of Agriculture (USDA) Agricultural Research Service (ARS). We also share practical advice to help organizations benefit from cloud computing for the life sciences.

### The Need for Long-Read Sequencing

Long-read sequencing allows progress on a large set of diverse, complex, and important scientific and technical issues. For example, resolving missing heritability linkages in population studies can enable better diagnosis and treatments for inherited diseases. Comprehensively characterizing the pathogens that underlie novel and antibiotic-resistant diseases can lead to the discovery and design of better vaccines, treatments, and outcomes.

## Table of Contents



In addition to its high value in human genomics, long-read sequencing is extremely beneficial in studying the complex genomes and transcriptomes of plants, insects, and livestock. With a complete view of their DNA and RNA, scientists are better able to improve selective breeding, develop natural growth enhancers, and precision-engineer genetic modifications. Advances in these areas can be essential to supplying food for a rising population on an increasingly fragile planet, or managing the next outbreak of a new infectious disease. By gaining a complete and comprehensive view of the genome of a given crop and its wild relatives, plant breeders can discover favorable gene alleles within a species itself, obviating the need to introduce foreign DNA via genetic modification.

But plant genomes and transcriptomes are so complex to sequence that we still lack a reference model for something seemingly as basic as bread wheat. In reality, bread wheat is an extremely complex genome to assemble—it's more than five times the size of the human genome, and it has three complete genomes in the nucleus of each cell, each consisting of around 3.5 billion base pairs.[1]

And, of course, sequencing an individual plant or animal is just a start. With the environment changing rapidly, there's greater need to understand how environmental changes are affecting gene expression. Scientists are also working to identify multiple genes that can affect a trait and understand the interplay among them. Factors such as yield, stress tolerance, and drought tolerance are complex traits that require examining whole genome interactions, other types of structural variation, epigenetics, plant-microbe interactions, and more.

Long-read sequencing targets many of the aspects of complex genomes and transcriptomes that make them difficult or impossible for short-read sequencing methodologies to handle:

- **Size and scale.** Plant genomes are many times larger than the human genome.
- **Repetitive regions.** Many plant and animal genomes have numerous repetitive regions, requiring detailed analysis and comparison to accurately determine where a DNA fragment belongs.
- **Polyploidy.** Many have multiple sets of homologous chromosomes, adding complexity.
- **Prevalence of non-model genomes.** The lack of well-studied reference models means researchers are often working "from scratch," conducting *de novo* assemblies with few or no guideposts.
- **Rising volumes of data.** Greater genome complexity increases the volumes of sequencer output that must be processed to align the fragments and join them into contiguous, error-corrected assemblies.

## Pacific Biosciences: Leadership in Long-Read Sequencing

Advances in long-read sequencing and associated analysis tools are making it possible to generate high-quality reference genomes for organisms that were too difficult to sequence using only short-read sequencing. This is especially true of the large, complex genomes that characterize many significant crop species.

Pacific Biosciences is a leader in long-read sequencing. The company was established in 2000 and is now publicly held. PacBio's Single Molecule, Real-Time (SMRT®) DNA Sequencing provides both very high consensus accuracy (capable of exceeding 99.999 percent accuracy when used with appropriate bioinformatics tools) and

exceptionally long read lengths (tens of thousands of base pairs long, compared to the hundreds typically produced by short-read sequencers). These longer reads reduce the overall number of DNA fragments that must be pieced together for *de novo* assembly, allowing greater clarity on repetitive areas and helping scientists create a frame on which to "hang" smaller fragments.

PacBio SMRT Sequencing has proven to be well suited for *de novo* assembly of larger or more complex genomes and transcriptomes, as well as tasks such as characterizing genetic variation and analyzing the role of methylation. PacBio long reads support *de novo* assembly both alone and in combination with short-read data to create hybrid assemblies. In addition to helping fill gaps inside scaffolds and joining contiguous regions of complex genomes, PacBio long reads are valuable for whole-genome sequencing of microbial organisms. The PacBio Sequencer is being widely used for both bacterial artificial chromosome (BAC by BAC) mapping methods and newer whole genome shotgun (WGS) sequencing approaches.[2]

The PacBio Sequencer combines high-performance optics, automated liquid handling, and an environmental control center, all directed through an intuitive touchscreen interface. Throughout the DNA sequencing process, the PacBio Sequencer delivers the recorded signals in real time to the primary analysis pipeline, facilitating high-throughput processing. Because the internal computational resources responsible for primary data analysis are integrated into the solution, chemistry and software advances can be seamlessly incorporated to provide immediate performance enhancements.

Through its SMRT Analysis package, PacBio offers a full suite of tools to analyze single-molecule sequencing data. SMRT Analysis tools support a broad range of applications including:

- *De novo* assembly
- Genome finishing and scaffolding
- DNA base modification detection
- Bacterial methylome and motif analysis
- Minor variant detection
- Compound mutations and phasing of distant SNPs
- Highly accurate consensus calling with variant detection

## Nimbix: Cloud Computing for the Post-Sequencer Pipeline

In addition to a sequencer's internal computational resources, genome assembly involves a pipeline of post-sequencer steps that are compute and data intensive. Many workloads that comprise the pipeline are highly parallel, allowing them to operate efficiently in a distributed, high-performance computing cluster or cloud environment.

By moving all or part of their processing pipeline to a secure, high-performance cloud, life science organizations gain flexible, expandable, and affordable access to the latest computing resources. IT groups can conserve space and power in their on-site data centers while avoiding the costs of acquiring and managing in-house HPC infrastructure. Researchers can meet their capacity and performance needs more quickly and economically, which can enable them to increase their research productivity, tackle larger data sets and more difficult problems, produce more accurate results by using more rigorous algorithms, and publish more quickly. With enterprise-focused cloud services, organizations can rapidly scale jobs up or down as needed, and be charged only for resources used.

"With most short-read data, you can develop in-house sequencing analysis pipelines using software like DNAStar* and Blast2GO*, or you can use public services like iPlant*. However, for larger data sets such as our 50x PacBio data, we need cloud computing with established pipelines, and a vendor with bioinformatics expertise. We would like to have absolute accuracy, but there isn't enough time, money, or manpower. Cloud gives you an alternative. We have pretty good infrastructure here for the next few years, so the cloud supplements our resources. But if someone is new to this, I would advise them to just go to the cloud. Cloud seems to be the future of bioinformatics."

– Brenda Oppert,
Research Molecular Biologist,
Stored Product Insect and Engineering Research Unit,
Agricultural Research Service Center for Grain and Animal Health Research,
US Department of Agriculture

Nimbix, Inc., is a cloud innovator specializing in HPC cloud services for technical computing in fields such as bioinformatics, energy, manufacturing, big data analytics, financial services, and media. In contrast to commodity web services platforms, the Nimbix Cloud is purpose-built to provide volume, speed, and simplicity for the most demanding HPC workloads, and its staff includes HPC subject-matter experts.

Nimbix's JARVICE supercomputer architecture offers LS organizations advanced controls over their HPC resources while providing the performance, ease-of-use, self-service, and elasticity that make cloud computing so powerful. Users can decide which system components and applications run on in-house equipment, and how much of the Nimbix public cloud to use for additional scale.

Using bare-metal resources and Intel® technologies, the Nimbix Cloud is optimized for bioinformatics processing tasks such as genome analysis, yielding results that can be up to 15 times faster than other cloud solutions.[3] To ensure enterprise-grade security, data is encrypted as it enters and exits the cloud, and workloads operate in containerized environments.

## Advancing Science at the USDA with Long-Read Sequencing

Genomic sequencing is central to the work of the USDA and its primary scientific, in-house research agency, the Agricultural Research Service (ARS).[4] ARS has locations in every state and a budget of USD1.1 billion for the 2015 fiscal year. Citing the growing importance of data-intensive computing, ARS is exploring next-generation approaches to obtain the advanced computing resources its scientists need to continue performing their mission-critical research.[5]

Numerous projects at ARS are using PacBio sequencers. For example, at the March 2015 Advances in Genome Biology and Technology (AGBT) conference, Dr. Timothy P. Smith of the ARS Genetics Breeding and Animal Health Research Center presented a goat assembly produced with PacBio sequence data. While a previous goat assembly generated from short-read data had a contig N50 of about 18 kb with hundreds of thousands of contigs, the PacBio assembly had a contig N50 of about 2.6 Mb and just 5,902 contigs. Dr. Smith's team is following up the goat effort with new projects to sequence the pig, sheep, and cow genomes using PacBio sequence data.[6] The team is also working to annotate functional aspects of the bovine genome, as well as to identify single nucleotide polymorphism (SNP) markers that can help breed heavier, disease-resistant cattle that mature more rapidly.

Dr. Brenda Oppert, a research molecular biologist, is using PacBio long-read sequencing and the Nimbix Cloud in her work at the Stored Product Insect and Engineering Research Unit of the ARS Center for Grain and Animal Health Research in Manhattan, Kansas. Dr. Oppert helped annotate the genome of the first sequenced beetle—the red flour beetle. She now leads a team that's using the PacBio Sequencer to target a second beetle: the lesser grain borer, *Rhyzopertha dominica*. A major pest of stored grains, *R. dominica* is difficult to control because the immature stages feed within the grain before the damage can be spotted, and fumigation-resistant strains are becoming a major problem at grain storage warehouses and processing facilities.

"We want to find better control methods based on a better understanding of the genome as well as the biochemistry, the physiology, the knowledge of how it survives, how the gut interfaces with the environment—many different approaches," Oppert explains. "We're working to identify vulnerabilities in the insect to help develop a new insect control that's specific to that pest and will have the least impact for mammals and beneficial insects." The team recently assembled 56x coverage of long-read PacBio data on the lesser grain borer genome using the Nimbix cloud.

Much insect research involves non-model genomes, making long-read sequencing assembly critically important. "There's extensive genetic data if you have models," Oppert explains. "Since we work with non-models, long-read data is absolutely necessary to get accurate assemblies." The team is using the Min-Hash Alignment Process with the Celera Assembler (MHAP/CA[7]) to assemble the *R. dominica* genome.

Once *R. dominica* is sequenced, the research team will continue to update, improve, and refine the genome. They will also explore new genomes, using long-read technology to sequence one insect of interest to stored products researchers each year, including insects that are larger and/or more complex.

Oppert's team also has several transcriptome projects at various stages of analysis, and sees the potential for long-read technology to accelerate that work. "We hope long-read sequencing can help with studying splice variants through RNA-Seq, although we haven't yet used it for those types of analyses," Oppert says. "The capability is there with long-read transcriptomes. The longer the read, the better idea you have of the mRNA."

Oppert has used nearly a half-dozen sequencing platforms in her career at ARS, and says each has advantages and disadvantages. "Cost often becomes the limiting factor," she says. "But when you factor in success, long-read technology is essential when you do not have a reference genome."

## Succeeding in the Cloud

Cloud computing provides a practical, affordable path to processing long-read data, according to Dr. Oppert. "With most short-read data, you can develop in-house sequencing analysis pipelines using software like DNAStar* and Blast2GO*, or you can use public services like iPlant*," she says. "However, for larger data sets such as our 50x PacBio data, we need cloud computing with established pipelines, and a vendor with bioinformatics expertise. We would like to have absolute accuracy, but there isn't enough time, money, or manpower. Cloud gives you an alternative. We have pretty good infrastructure here for the next few years, so the cloud supplements our resources. But if someone is new to this, I would advise them to just go to the cloud. Cloud seems to be the future of bioinformatics."

Given the specialized requirements of genome sequencing, Oppert believes it's important for life science users to choose a cloud service provider with expertise in technical computing. This expertise can be essential to ensuring their cloud infrastructure provides optimal performance and reliability for life science workloads. It can also help organizations keep pace with the latest advances in algorithms and applications.

"Most of us in the life sciences don't have the extensive programming background we would need to do this on our own," Oppert says. "It's important to find a cloud provider that has that expertise and can work with you on your specific projects and help you stay within your budget. Make sure they support the particular algorithms that are important to you. The algorithms will keep advancing, though, so make sure the cloud provider is keeping up with the technology. This is a fast-moving train. It can be challenging to keep up."

## Intel's Role: Advancing Technologies and Solutions for Life Sciences Computing

Intel technologies are the foundation for powerful, flexible HPC cloud services. The Intel® Xeon® processor E7 and E5 families enable faster results and greater accuracy by providing significant improvements in performance, power efficiency, virtualization, and security. The newest Intel Xeon processors support a software-defined infrastructure, allowing greater flexibility with higher levels of automation and orchestration for cloud environments. Intel Xeon processors also offer hardware-assisted support for virtualization performance and security in cloud environments.**

The Intel® Xeon Phi™ coprocessor, based on Intel® Many Integrated Core™ technology, enables LS users and developers to increase performance for highly parallel workloads without limiting flexibility or investing the time typically needed for application-specific integrated circuit (ASIC) and general-purpose graphics processing unit (GPGPU) technologies.

In addition to technology building blocks, Intel collaborates with the open source community and other software developers to expand the universe of optimized applications and algorithms for life sciences computing, including genomic data processing. Intel's research and development activities drive advances in server, storage, and networking technologies to make cloud computing more powerful, reliable, and secure. Beyond our own R&D activities, Intel Science and Technology Centers in the United States and Intel Collaborative Research Institutes internationally are Intel-funded, jointly-led research collaborations between Intel and the academic community, conducting R&D that advances tomorrow's cloud computing solutions.

"The level of change that we're seeing in the life sciences and genomics is going to change our lives in every aspect. Precision medicine is happening right now. It's going to change the way we practice medicine and diagnose and treat illness. And we're moving toward precision agriculture, where we'll be able to design very sophisticated interventions that will have a specific target and will minimize other impacts. It's exciting."

– Brenda Oppert,
Research Molecular Biologist,
Stored Product Insect and Engineering Research Unit,
Agricultural Research Service Center for Grain and Animal Health Research,
US Department of Agriculture

Intel also advances progress through investments in technology startups. Since 1991, Intel has invested more than USD11.4 billion in over 1,400 companies in 57 countries—including Pacific Biosciences. In that time, more than 211 portfolio companies have gone public on various exchanges around the world and more than 369 were acquired or participated in a merger.

## Continuing Challenges and Benefits for Life Sciences Computing

The need for long-read sequencing and HPC resources will continue to grow—but so will the benefits and the potential importance. Ongoing advances in long-read sequencing and post-processing—driven by industry innovators such as Intel, Pacific Biosciences, and Nimbix—will support efforts to adjust to a changing planet, feed a global population that is predicted to reach nine billion by 2050, and deliver personalized medical therapies.

"The level of change that we're seeing in the life sciences and genomics is going to change our lives in every aspect," says Dr. Oppert. "Precision medicine is happening right now. It's going to change the way we practice medicine and diagnose and treat illness. And we're moving toward precision agriculture, where we'll be able to design very sophisticated interventions that will have a specific target and will minimize other impacts. It's exciting."

## Move Forward Together

How is your organization using long-read sequencing and cloud computing? How can Intel, PacBio, and Nimbix support your success? To learn more:

Talk to your Intel representative, or visit us on the web: www.intel.com/healthcare/bigdata, www.intel.com/healthcare/optimizecode, www.pacificbiosciences.com or www.nimbix.net

Follow us on Twitter:

- @portlandketan, @IntelHealth
- @infoecho, @PacBio
- @stevemhebert, @Nimbix
- @USDA

Join the Intel Health & Life Sciences Community for ongoing updates, discussions, and more.




[1] WheatGenome.info: An integrated database and portal for wheat genome information. Kaitao Lai, Paul J Berkman, Michal Tadeusz Lorenc, Christopher Duran, Lars Smits, Sahana Manoli, Jiri Stiller, David Edwards. Plant and Cell Physiology (2012) 53(2): e2.

[2] For published papers, see http://www.pacificbiosciences.com/news_and_events/publications/

[3] For example, see Steve Hebert, High Speed BWA in the Nimbix Cloud, July 25, 2012. http://www.nimbix.net/blog/2012/07/25/high-speed-bwa-in-the-nimbix-cloud/

[4] Please note: Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture or the Agricultural Research Service.

[5] See Big Data and Computing: Building a Vision for ARS Information Management. Workshop Summary. Feb. 5-6, 2013. USDA Agricultural Research Service. http://www.ars.usda.gov/SP2UserFiles/Place/20800500/BigDataReport_Mar-7-2013.pdf

[6] For a blog entry describing some of Dr. Smith's earlier work with PacBio sequencers, see PacBio Blog, AGBT Day 2 Highlight: Bovine Immune Response Characterized with PacBio Sequencing, Feb. 22, 2013. http://blog.pacificbiosciences.com/2013/02/agbt-day-2-highlight-bovine-immune.html

[7] For information on MHAP, see Konstantin Berlin et al., Assembling large genomes with single molecule sequencing and locality-sensitive hashing, 33, 623-630 (2015) doi: 10.1038/nbt.3238, http://www.nature.com/nbt/journal/v33/n6/full/nbt.3238.html. And Eugene W. Myers et al., A Whole-Genome Assembly of Drosophila. Science 287 2196-2204.



Printed in USA 0715/KP/TDA/XX/PDF ♻ Please Recycle 332723-001US