WHITE PAPER

Bioinformatics

BioNano Genomics® and Intel® Xeon Phi™ Coprocessor Advance Next-Generation Mapping

With the Intel Xeon Phi coprocessor to speed throughput, BioNano Genomics' Irys® System helps researchers improve the quality and speed of genome finishing and structural variation analysis

Ketan Paranjape
General Manager,
Health and Life Sciences,
Intel Corporation

Mark Borodkin
Vice President,
Systems Development and Operations,
BioNano Genomics

Overview: Putting Genome Maps Back in the Spotlight

Genome maps were at the center of early genomic research—vital tools that let researchers see the locations of genetic markers across large segments of a genome and make sense of their structure. But the high costs and months of effort needed to produce these maps put them beyond the reach of all but the best-funded efforts. The industry evolved, and next-generation sequencing (NGS) became the dominant tool for genomic research.

But even as NGS read-lengths have grown longer, the structural complexity of genomes forces researchers to put forth massive amounts of time and effort to assemble an accurate and complete genome from the fragmentary output of NGS sequencers. This results in genome assemblies that still miss much of the information that sequencing data is unable to decipher.

BioNano Genomics is driving the next great evolution in genomic analysis techniques—and bringing genome maps back to the fore. BioNano Genomics' Irys® System provides optical next-generation mapping (NGM) that dramatically reduces the cost and complexity of traditional genome mapping. In doing so, BioNano Genomics is enabling researchers to generate the highest quality genome assemblies available on the market today, as well as perform stand-alone structural variation (SV) analysis without the use of NGS.

To deliver outstanding throughput and scalability for its integrated IrysSolve® analytics pipeline, BioNano Genomics developed IrysSolve® Compute, based on the Intel® Xeon Phi™ coprocessor. Taking advantage of the coprocessor's massively parallel processing capabilities, the platform is helping customers accelerate time-to-results and generate new insights into plant, animal, and human genomics. In production work performed at BioNano Genomics' in-house lab, the workload acceleration enabled by the coprocessors reduced the time needed for human genome assembly from 108 hours on a general-purpose server to less than 24 hours on IrysSolve Compute using the Intel Xeon Phi coprocessor—more than a fourfold improvement.

Table of Contents

Innovation to Transform the Map

The key to generating useful genome maps is precise, long-range detection of labels across thousands of bases, up to whole chromosomes. One of the major challenges is that DNA is an inherently flexible molecule that naturally forms coils in solution, hindering accurate, direct measurement.

BioNano Genomics addresses these and other challenges through the Irys System, a comprehensive solution that delivers breakthrough capabilities in single-molecule, nanoscale, whole genome analysis technology (Figure 1).

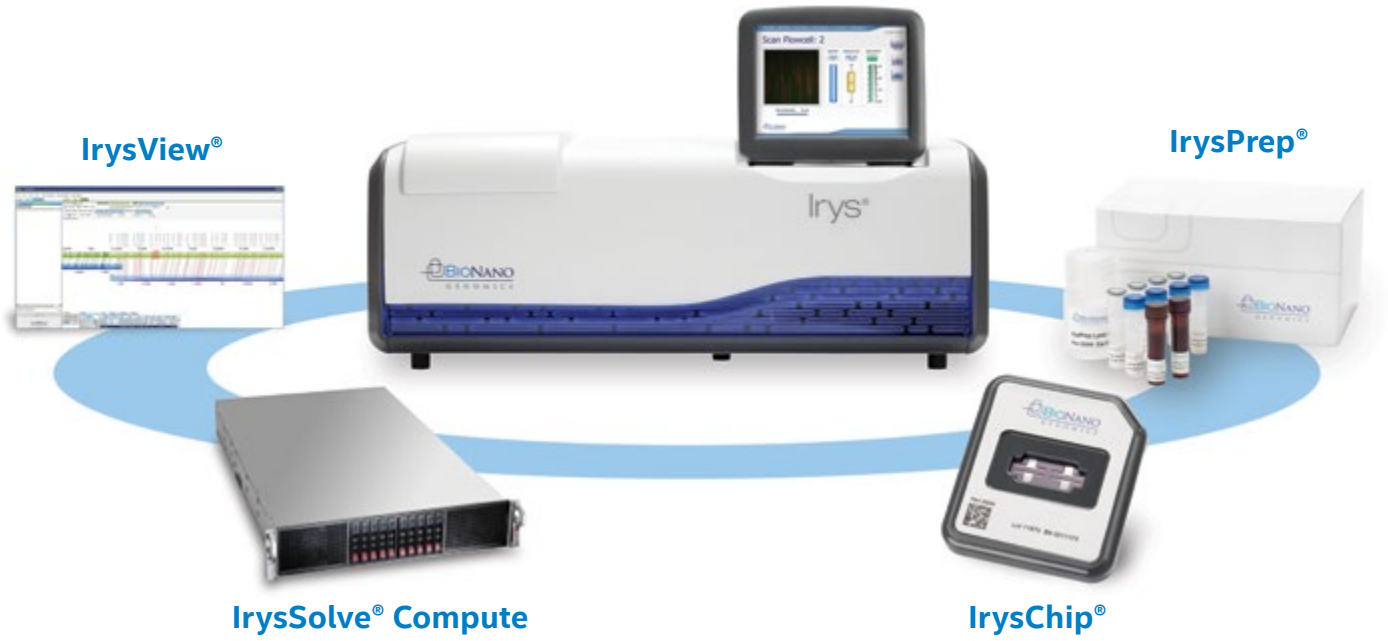

Figure 1. BioNano Genomics Irys System

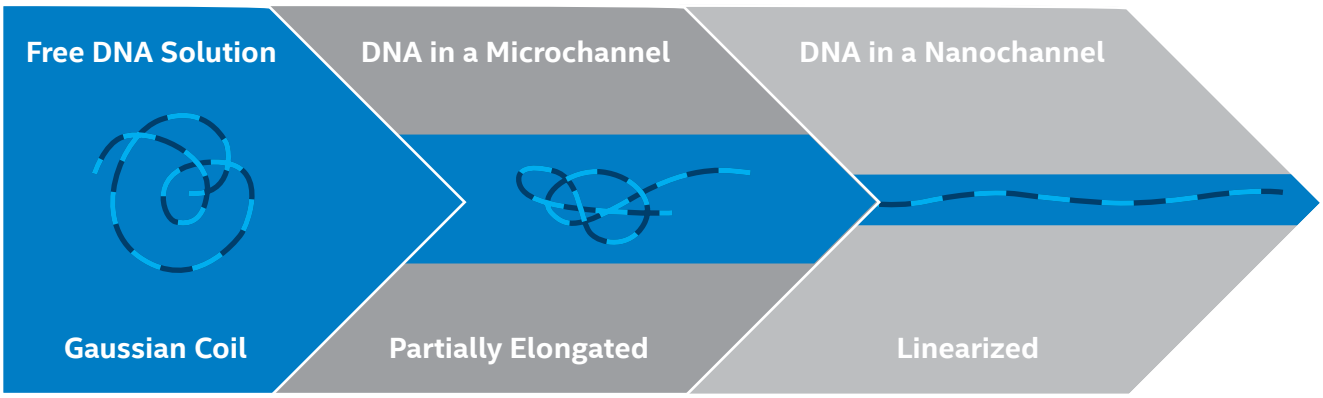

Figure 2. Intact, Linearized DNA

BioNano Genomics was founded in 2003 and has its headquarters in the booming life sciences center of San Diego, Calif. Led by a team of experts in the field, BioNano Genomics developed innovations that use optical mapping and massively parallel NanoChannel arrays on a proprietary chip to unravel, linearize, image, and map long, genomic DNA fragments at rates of up to several gigabases of DNA per hour (Figure 2).

In contrast to methods that require front-end amplification or shearing DNA into small fragments, BioNano Genomics' Irys System processes native-state DNA. The system keeps very long sequences of double-stranded DNA intact, preserving valuable structural information.

Unamplified DNA molecules are labeled using IrysPrep® reagents, and then electrophoretically guided through NanoChannel arrays on the IrysChip®. The system uses laser excitation to illuminate the fluorescently labeled DNA on the NanoChannel arrays. An onboard CCD camera, coupled with a proprietary auto-focusing mechanism and control software, rapidly takes pictures of these linearized DNA molecules.

Molecules move through thousands of parallel channels simultaneously, enabling high-throughput processing for a more accurate genome map. Each NanoChannel ensures a uniform DNA stretch and prevents molecules from folding or overlapping other molecules, producing high-resolution imaging of single molecules. This provides the ability to accomplish *de novo* assembly (without the use of a reference) of large genomes for which it is impractical and cost-prohibitive with other mapping platforms.

Prior to analysis, the raw images of labeled long DNA molecules are converted to digital representations of the motif-specific label pattern. These patterns are then assembled *de novo* to create the genome consensus map of the original genome. The Irys System scans molecules that are up to 2.5 megabase long, and creates a comprehensive genome map from as little as 160 ng of labeled DNA loaded onto an IrysChip.

Organizations are using the Irys System and next-generation mapping for the following uses:

- **Assembly validation.** Users compare NGS contigs or scaffolds against a genome map to gain independent validation of their assemblies and identify any misassemblies.
- **Genome finishing using the Irys System's hybrid scaffolding capability.** Users import NGS contigs to order and orient them, using the genome map to provide essential scaffolding information, define gaps, and build hybrid scaffolds. This process can be iterated to generate increasingly accurate and complete genomes.
- **SV analysis.** Users compare the genome maps to reference sequences to identify structural variants, which they can then display on the IrysView visualization software or a third-party genome browser. This capability is particularly valuable in identifying the genomic contributions of mothers and fathers to difficult-to-diagnose diseases or to discover novel *de novo* variants of clinical relevance.
- **Comparative genomics.** Users visualize and examine variation between two genome maps to find structural variation or differences in gene content and architecture.

Breakthroughs enabled by BioNano Genomics and its customers will help make Intel's vision of All in One Day precision medicine a reality.

Accelerating the Analytics Pipeline with the Intel® Xeon Phi™ Coprocessor

BioNano Genomics' IrysSolve software runs custom algorithms that use modern multithreading and parallelization techniques to process the unique Irys data type. Data generated by the Irys System can be exported at any step in the pipeline to support custom analysis or be used with other applications. The solution's Linux* implementation allows for easy interfacing with other technologies and workloads.

IrysSolve software can run on workstations, servers, and general-purpose clusters, but with the rapid growth in the demand for genomic analysis, BioNano Genomics wanted to offer an optimized hardware platform that could help customers keep pace with the Irys instrument's high-volume outputs. The resulting IrysSolve Compute solution uses the Intel Xeon Phi coprocessor to dramatically increase analytics throughput.

The IrysSolve Compute solution is a compact, affordable, and high-performance cluster powered by six first-generation Intel Xeon Phi coprocessors and built using a Super Micro Computer server. Intel Xeon Phi coprocessors are based on the massively parallel Intel® Many Integrated Core (Intel® MIC) architecture. The coprocessors are compatible with the widely used Intel® Xeon™ processor programming model and designed to deliver maximum performance for thread-parallel and data-parallel workloads.

To gauge the performance improvements of the Intel Xeon Phi coprocessors, BioNano Genomics ran a human genome assembly that took 108 hours on its previous system—a server powered by two Intel® Xeon® processor E5-2680 v3 and supporting 48 hyperthreads on 24 cores. The Intel Xeon Phi coprocessors used in the tests have 60 cores apiece, with each core able to support four hyperthreads, for a total of 1,440 threads on the 360 cores of the six-coprocessor system. The dramatic increase in parallel processing capability enabled the IrysSolve Compute solution to complete the job in just under 24 hours.

Having an Impact

BioNano Genomics and its customers have mapped more than 200 unique species and more than 300 unique humans. Published papers have highlighted significant results. For example:

- A paper in *Nature Methods* reported that researchers led by Pui-Yan Kwok achieved high-quality *de novo* human genome assemblies with unprecedented contiguity by combining the Irys® System's NGM with Illumina-based approaches for NGS. The research team demonstrated that when using the GemCode* technology from 10X Genomics sequencing library preparation, the resulting hybrid assembly had comparable or slightly better contiguity than that of NGM-NGS hybrid assemblies achieved with Pacific Biosciences in previous studies.[1]
- Mak *et. al.* reported in *Genetics* that using the Irys System for stand-alone SV detection of phase 1 data from the 1000 Genomes Consortium pilot, they were able to detect and validate seven times more SVs (large insertions and deletions, in this case) than were previously found in the data.[2]

BioNano Genomics uses the IrysSolve Compute solution to handle workloads in its own high-throughput lab in San Diego. The company runs mapping projects in collaboration with academic and industry partners, as well as to show potential customers how genome mapping can address their specific research challenges.

Among recent projects, BioNano Genomics is supporting efforts to complete a high-quality assembly of the genome of the mosquito that carries the Zika virus. The team needed just one day to generate data on the Irys System and another day to analyze it on the Intel® Xeon Phi™ coprocessor-based platform and share it back with the Zika consortium. The assembly team expects to reiterate the hybrid assembly process as the NGS teams further refine the sequence.

User Profiles: Resources to Improve Plant, Animal, and Human Genomics

Organizations are using the IrysSolve Compute solution with its Intel Xeon Phi coprocessors in diverse research that may help feed a rising population, create more hospitable habitats for beneficial insects, and inspire new ways of diagnosing, treating, and preventing disease. Researchers are enthusiastic about next-generation mapping, including the performance and convenience of the IrysSolve Compute solution. They also speak highly of BioNano Genomics as a committed collaborator and innovator.

Kansas State University: Seeing the Bigger Picture

At the Kansas State University (KSU) Bioinformatics Center, the Irys System and Intel Xeon Phi coprocessors are strengthening biomedical research at the university and beyond. The center was established in 2001 and is funded by the National Institutes of Health through the Kansas Institutional Development Award (IDeA) Network of Biomedical Research Excellence (K-INBRE).

Dr. Susan Brown, Distinguished Professor at KSU, has directed the Bioinformatics Center since she helped establish it in 2001. She's been involved in genome research since the 1990s and was eager to exploit the opportunities presented by BioNano Genomics' affordable genome mapping. KSU started working with BioNano Genomics before the company offered a commercial product. The center became a beta user in Nov 2012, and purchased its first Irys System in April 2013. It has since added a second Irys System and an IrysSolve Compute solution, and will add a second IrysSolve Compute solution shortly.

KSU researchers are using the BioNano Genomics platforms primarily for research focused on insect biology, while others are applying the platforms to their studies of humans, chimps, alpacas, electric eels, chickens, crows, "lots and lots of plants," and more. Researchers use the platforms both to supplement their NGS-based *de novo* assemblies and to perform SV discovery.

"BioNano Genomics gives us better contiguity than NGS alone," Brown explains. "It helps us fill gaps in our knowledge and understand the architectural arrangement of these large, complex genomes. In addition, because the Irys is a completely independent platform, it can validate the assembly of the sequence data."

The impact on SV discovery is especially promising for cancer research. "The power of BioNano Genomics' approach and the need for dedicated servers with the Intel Xeon Phi coprocessor is the ability to do comparisons between experimental and control or diseased and normal genomes at the level of optical mapping, to complement the things you can do at the nucleotide level," says Brown. "It gives you the bigger picture, so you can get at whether the changes in a genome are due to copy number repeats or large structural variations. If you work in medicine, you can compare my genome to my cancer genome, instead of comparing it to some reference genome. That's really cool, because it helps us get more directly to the cause of a given mutation. With some cancers, SVs are more closely associated with those mutations, so being able to get at SVs directly is really helpful."

From a Week to a Day

The Bioinformatics Center uses the state's largest supercomputer, KSU's Beocat* cluster, for its NGS assembly and analysis work. But the center processes all outputs from the Irys instrument on IrysSolve Compute using the Intel Xeon Phi coprocessor.

"The bottleneck for this kind of work is the bioinformatics, not the sample prep," Brown says. "The Intel Xeon Phi coprocessor has totally sped up the computations, and that's a major advantage. The dedicated system is helpful because it's right here, and it's completely under our control. When we were doing our mapping on the cluster, it took time to get in the queue, and you never knew if you'd asked for enough memory and so forth. It often took the better part of a week to get results. Now, we get them overnight."

The mapping results are empowering for scientists. "A lot of people who are interested in optical mapping have worked hard to get their genome and improve the contiguity of their genome and the assembly," Brown comments. "It's frustrating for them to still have megabase scaffolds that are not anchored to a chromosome. So it's fun to see the data coming off the scaffolding software and to see that you've been able to put together several more contigs and to map scaffolds into the chromosome. That in itself really moves people's research forward. It's very satisfying."

Now, the scientific insights are coming more quickly. "It used to take whole laboratories, and two or three NIH grants, and a year of effort to develop a restriction map of a complex genome," says Brown. "When we first started using BioNano Genomics but without the Intel Xeon Phi coprocessors, we could do it in two or three months. Now it takes about two weeks. In many cases, the total turnaround time, from when we get the sample until we have the full assembly, is within a month. It's not a silver bullet, and there's still work to be done, but BioNano Genomics mapping is economical and efficient. It's a major, major improvement."

> "It's fun to see the data coming off the scaffolding software and to see that you've been able to put together several more contigs and to map scaffolds into the chromosome. That in itself really moves people's research forward. It's very satisfying."
>
> Dr. Susan Brown,
> Distinguished Professor,
> Kansas State University

VIB: Taming Monsters

VIB, a life sciences research institute in Flanders, Belgium, deployed an Irys System and IrysSolve Compute solution to support the institute's commitment to innovation for economic development and entrepreneurship. VIB is based on a close partnership with five Flemish universities—Ghent University, KU Leuven, University of Antwerp, Vrije Universiteit Brussel, and UHasselt—and is supported by a solid funding program. VIB's technology transfer activities translate basic research results into new economic ventures which, in time, lead to innovative products that can be used in medicine, agriculture, and other applications.

Because of its strong entrepreneurial focus, VIB actively searches for and investigates promising new technologies. "We have to be on the front line of technology to be making those breakthrough discoveries that will be welcome in the marketplace," says Dr. Stéphane Plaisance, Senior Scientist and Data Miner at the Leuven Nucleomics Core. "BioNano Genomics technology helps us be on the front line. It gives us an advantage in opening our eyes and putting us in the field early."

The Nucleomics Core is a service facility providing advanced scientific equipment and expertise for VIB's internal users, collaborators, and external customers. Dr. Plaisance is a biologist whose work at the Core includes advising researchers on how to take full advantage of those resources. He says the Irys System and IrysSolve Compute solution provide breakthrough capabilities for researchers and makes genome mapping possible for a broader range of research organizations.

"Before BioNano Genomics, it was almost impossible for a non-expert group to do full whole genome sequencing of large genomes," he says. "You could do a good assembly for a genome of less than 1 gigabase, but beyond that, you would very often miss pieces of the assembly or fit them together incorrectly. BioNano Genomics has become affordable, and it lets you look at the structure of the genome as well as puts together the pieces of sequence obtained from the assembly."

Better Quality Genomes Faster

The Nucleomics Core provides BioNano Genomics services to VIB users, as well as to external research groups, and biotech companies working to characterize living organisms or to improve commercial crops. The Irys System and IrysSolve Compute using the Intel Xeon Phi coprocessor serves as a finishing tool to improve the quality of NGS assemblies based on the optical map data. "Our customers have been struggling to get very high-quality assemblies because their model organisms are often polyploid and as complex as some kinds of cancers, with multiple duplications in their history. They're monsters," Plaisance says.

> "BioNano Genomics lets you push your genome assembly one step further. You get a second, independent opinion, and it gives you a nice cartography of your genome... It reduces the complexity of the problem and allows you to achieve a better genome build in a fraction of the time."
>
> Dr. Stéphane Plaisance,
> Senior Scientist and Data Miner,
> Leuven Nucleomics Core (VIB)

BioNano Genomics' genome mapping with the Intel Xeon Phi coprocessor helps these researchers get a more accurate genome map more quickly. "BioNano Genomics lets you push your genome assembly one step further," says Plaisance. "You get a second, independent opinion, and it gives you a nice cartography of your genome. In addition, BioNano Genomics genomic maps are larger than NGS scaffolds and are not biased by the computing process. There are still gaps to fill, but BioNano Genomics reduces the number of pieces by a factor of up to 10, and it gives you some guidance to correct and fill the gaps. It reduces the complexity of the problem and allows you to achieve a better genome build in a fraction of the time."

The throughput delivered by the Intel Xeon Phi coprocessor means Plaisance and the BioNano Genomics technology adopters get assembly results in a few days of computing rather than in several weeks. This encourages scientific creativity and discovery.

"IrysSolve Compute using the Intel Xeon Phi coprocessor has strongly facilitated our work," Plaisance says. "The difference is day and night. The ease of mapping and the speed improvement are so tremendous that things we couldn't dream of before we can now do easily, like repeating a run with different parameters. They have relieved the main limitation in the primary analysis of the optical data. With the Intel Xeon Phi coprocessor, we can build an optical assembly even for human and plant-sized genomes within one to a few days. It is hugely accelerating our throughput."

French National Institute for Agricultural Research: Complementing Long-Read Sequencing

For large, complex genomes, even long-range sequencing platforms often leave gaps and ambiguities that lead to inaccurate or incomplete assemblies. At France's Plant Genomic Resources Center (CNRGV), part of the French National Institute for Agricultural Research (INRA), researchers and collaborators recently acquired an Irys System and IrysSolve Compute solution to complement their use of Pacific Biosciences' Single Molecule, Real-Time (SMRT*) long-read sequencing systems.

INRA is Europe's top agricultural research institute and a leading center for the agricultural sciences. INRA established CNRGV in 2004 as a genomic repository center dedicated to assisting plant genomic programs. Located in Toulouse, the center provides resources and services to French, European, and international users, working with a number of labs to accelerate both basic research and projects with practical applications. Many projects involve critical work in areas such as enhanced plant breeding and pest control.

Complex Genomes

INRA's research partners are studying large, complex plant and animal genomes, many of which have a high number of repetitive elements. Researchers are often studying particular regions of a genome that are associated with specific traits. And they're eager to bring the power of genome mapping to their research challenges.

"Our partners are completely excited by the genome mapping capabilities of the BioNano Genomics system," says Dr. Hélène Bergès, Managing Director of the CNRGV. "Many of our collaborators want to test it. Even with high-quality sequencing equipment and improvements in read lengths, there is still missing data and misassembled data. BioNano Genomics lets researchers improve the quality of the data, and get more data on the alignment and the organization of the contigs. Researchers can increase the size of the scaffolds, and obtain a more accurate, reliable assembly. They also get independent confirmation of their data."

Project teams are using or planning to use the Irys platforms to map the sunflower, passion fruit, oak, corn, maize, and ladybird genomes, among others.

Efficient and Integrated

CNRVG chose IrysSolve Compute using the Intel Xeon Phi coprocessor to gain an efficient, integrated solution that would provide analytic results more quickly than a general-purpose cluster. "We want answers as quickly as possible," says Dr. Céline Chantry-Darmon, research engineer at CNRGV. "In addition, we work with a great diversity of customers, so we need a ready-to-use system with reproducibility. The integrated system is simpler for them to use and well suited to the diversity of the projects we work on."

BioNano Genomics earns praise for helping users take full advantage of their new tools. "The BioNano Genomics support organization has done an excellent job in transferring their expertise to us, both when it came to installing the system but also continuously during our first projects," says Bergès. "If we have any questions, they are always there to help us."

Scientists at CNRGV see the BioNano Genomics platforms as important tools to advance the entire field of genome studies and support emerging multidisciplinary approaches. "The field is producing more and more data, but there are big problems with the reliability of the data in the public databases," says Bergès. "With BioNano Genomics, we are better able to determine if the assembly in the database is correct. If we can confirm the reliability and completeness of the data, and connect the genetics, genomics, epigenetics, phenotypic data, and other approaches, then we will have something more useful. When we can link all the different approaches to create knowledge, then we will have a greater understanding of the plants and their diversity by taking advantages of the possibility to study structural variation with the BioNano Genomics platforms. That is why BioNano Genomics is so important."

> "The field is producing more and more data, but there are big problems with the reliability of the data in the public databases. With BioNano Genomics, we are better able to determine if the assembly in the database is correct."
>
> Dr. Hélène Bergès,
> Managing Director,
> CNRGV (INRA)

Expanding Toward Clinical Workflows and All in One Day Precision Medicine

BioNano Genomics is leading a wave of innovation that meshes with Intel's vision of All in One Day precision medicine. Intel envisions a world where cancer patients receive precision treatment plans based on their unique biomolecular profile within 24 hours of diagnosis. Intel is driving innovation on many fronts and collaborating with health, life science, government, and technology leaders to make All in One Day a reality. The breakthroughs enabled by BioNano Genomics and its customers will be crucial in this effort.

Moving forward, BioNano Genomics expects to further expand its capabilities in the discovery of clinically useful heterozygous variants, with an eye to putting its workflow into clinical environments. This expanded focus calls for even higher levels of scalability, performance, and throughput, and BioNano Genomics looks forward to tapping into the capabilities of the second-generation Intel Xeon Phi processor, formerly codenamed Knights Landing. This product family combines the performance of massively parallel offload accelerators with the benefits of a server-class processor, to increase programming efficiency, reduce costs, and deliver even faster time-to-answers.

Through ongoing innovation and collaboration, Intel and BioNano Genomics will continue to provide powerful, cost-effective solutions that bring the power of genome mapping to leading-edge research in plant, animal, and human genomics.

Learn More

Explore Intel solutions for big data in health and life sciences. Talk to your Intel representative, or visit us on the web.

www.intel.com/healthcare/bigdata
www.intel.com/healthcare/optimizecode

Learn more about BioNano Genomics

www.bionanogenomics.com

Stay informed

@portlandketan
@IntelHealth
@bionanogenomics

[1] See BioNano Next-Generation Mapping Yields Superior Genomic Assemblies with Greater Contiguity, http://bionanogenomics.com/wp-content/uploads/2016/05/Bionano-Nature-Methods-Irys_illumina-Press-Release.pdf., and Yulia Mostovoy, et.al., A Hybrid Approach for De Novo Human Genome Sequence Assembly and Phasing, Nature Methods, 13, 587–590 (2016) doi:10.1038/nmeth.3865. http://www.nature.com/nmeth/journal/vaop/ncurrent/full/nmeth.3865.html

[2] Angel C.Y. Mak, et.al., Genome-Wide Structural Variation Detection by Genome Mapping on NanoChannel Arrays, Genetics January 1, 2016 vol. 202 no. 1 351-362; DOI: 10.1534/genetics.115.183483. http://www.genetics.org/content/202/1/351.long