WHITE PAPER

Accelerating Plant and Animal Genomics for Biodiversity with the Latest Intel® Technologies

Smithsonian Institute for Biodiversity Genomics halves genome assembly time using Intel® Xeon® processor E7 v3 family and Intel® Solid-State Drives

Rebecca B. Dikow, Ph.D.
Postdoctoral Fellow
Smithsonian Institute
for Biodiversity Genomics

Sandeep Gupta
Platform Applications Engineering
Software Services Group
Intel

Mathew H. Taylor
Senior Solutions Strategist and Architect
Health & Life Sciences
Intel

Overview

As sequencing machines advance in capabilities and cost effectiveness, genomic analysis in scientific research is gaining momentum, expanding from human genomics to the study of other species on the planet. Historically, genomic assembly and analysis require massive computing resources—typically a high-performance computing (HPC) cluster having a large number of nodes and cores with sharable memory and storage. However, the logistics of managing and configuring such a cluster slows the pace of scientific research because the lengthy time periods needed to finish an individual experiment can limit the number of experiments that can be run in a given time window. Execution run-times can vary widely, depending on what resources are available for a particular job.

Researchers at the Smithsonian Institute for Biodiversity Genomics encountered these issues in using the Smithsonian Institution's shared HPC cluster, which consists of more than 80 servers with approximately 3,300 cores, 16 TB of RAM, and networked storage. The Smithsonian Institution is the world's largest museum and research complex, and the Institute for Biodiversity Genomics was established to accelerate genomic studies that can help humans understand and preserve biodiversity across the tree of life. Using a subset of the cluster, genome assemblies often took weeks to complete. Some large assemblies failed to run to completion, causing frustration for scientists and further slowing the research pipeline.

Smithsonian researchers saw Intel's latest data center technologies as an opportunity to meet the Institute's growing need for HPC capacity while reducing cost, complexity, and space requirements. They worked with Intel to explore the performance impact of these technologies—which offer higher core and thread counts, tremendous memory capacity, and high-capacity solid-state drives (SSDs)—for their plant and animal genomic studies.

Using a single server powered by the Intel® Xeon® processor E7-8890 v3, Smithsonian researchers assembled the genome of a zebrafish (Maylandia zebra) using publicly available data in 47 hours—approximately half the time needed on a subset of the cluster and weeks faster than it would have taken on state-of-the-art hardware just a few years ago. The new server also used dedicated solid-state drives, specifically the Intel® Solid-State Drive (Intel® SSD) Data Center (DC) Family for PCIe* P3700 series, to enhance storage performance.

Table of Contents

The team also demonstrated significant performance improvements on shorter test runs using subsets of the data. These shorter runs give scientists more flexibility and speed with their experiments, allowing them to test more hypotheses, think more creatively, and accelerate time-to-insights. The performance achievements also offer greater flexibility to work with genomes of different sizes, an important capability for plant and animal genomics.

Genome assembly is one crucial step in genomics research, and it has a ripple effect throughout the research pipeline. Considering the millions of species on the planet and the rapid rate at which they are going extinct, as well as their importance to human health and society, the impact of biodiversity research has the potential to be truly profound. But the amount of data to be generated, processed, and stored is staggering. Technology innovations that increase computing performance and capacity for genomic data in sustainable, energy-efficient ways will be critical to sustaining global biodiversity. These innovations will also be of exciting, long-range importance to medicine, energy, agriculture, and other fields.

High-Performance Computing for a Sustainable Planet

High-performance computing is central to genome assembly and analysis. The life sciences are among the biggest of big data users, and the field's data volumes and analytics requirements are rising rapidly. With a single genome sequencer churning out 3.6 TB of data in six days,[1] researchers predict the data requirements for genomics alone will equal or exceed the data requirements of YouTube* or the entire field of astronomy by 2025.[2]

While datasets are growing, they're only half the story. Data alone do not produce scientific results; analysis is required. Published genome assemblies and analyses have generally taken weeks or months to complete. To keep pace with the growing amount and importance of genomic data, it is imperative that computation scale with the datasets and analysis goals.

Although much attention has focused on human genome sequencing and the promise of precision medicine, genomic information about the plant and animal worlds is equally crucial. On a fragile, fast-changing planet, the ability to assemble and analyze the genomes of diverse species will be critical to the survival of a great many species and ecosystems. Biodiversity genomics, which seeks to gather, analyze, and compare the genomes of organisms across the tree of life, can help scientists understand threatened species and ecosystems, and improve efforts to preserve them. Knowledge of plant and animal genomes can increase our ability to manage climate change, feed a rising population, and mitigate the impact of newly emergent diseases.[3] Genome biology is a source of innovation through areas such as drug discovery, food safety, and healthcare, including stem cells, organ/neural regeneration, and aging.

Beyond direct human benefits, we're living in a critical time for documenting the diversity of species and genome biology on earth: humans have a limited time to capture and understand biodiversity before large-scale extinction makes it too late. Human activity is conservatively estimated to have eliminated approximately 900 species in the past 500 years. The International Union for the Conservation of Nature (IUCN) has approximately 17,000 plants and animals on its Red List of endangered species, noting that one in eight birds, one in four mammals, one in five invertebrates, one in three amphibians, and half of all turtles are in danger of extinction.[4]

Biodiversity Genomics Challenges

Biodiversity genomics has particular challenges that set it apart from the study of human genomics. Plant and animal genomes are extremely diverse. Something as seemingly simple as bread wheat has a genome that is more than five times larger than the human genome and is so difficult to assemble that we still lack a reference genome for it. While we know a great deal about the Drosophila and C. elegans (a nematode), there are millions of plant and animal species about which we know very little. This leaves scientists to work with non-model genomes, assembling genomes "from scratch," via de novo assembly. Often, because plant and animal genomes are less well-known and are different from the human genome, our analytical tools simply don't work as well.

Genome Size

Genome sizes across the tree of life vary greatly compared to human and most mammal genomes, which are all about three billion nucleotide base pairs (3 giga-base pairs or Gbp). Many birds are around 1 Gbp, but a sunflower genome is around 3 Gbp, and a Loblolly pine tree is 22 Gbp. There are also more outliers. One spider genome might be 1 Gbp, and another 8 Gbp. Genome size doesn't always correlate with the number of genes in a genome; it may have resulted from ancient genome duplications or just a large amount of repetitive DNA that does not code for any gene or protein. Larger genomes mean that more sequencing is required to "cover" the whole genome. Beyond the quantity of sequencing, highly repetitive and duplicated genomes are much harder to assemble, requiring vast computational resources.

Sample Availability

Sample availability is the biggest limitation to generating genome data across the tree of life. For many species that are tiny or live in inaccessible environments, DNA samples are difficult to obtain. Species that are extinct, or for which samples were collected long ago, present further challenges because DNA degrades quickly when not preserved properly. To successfully sequence a complete genome, scientists need good quality (freshly preserved) tissue with enough not-highly-fragmented DNA for analysis. DNA can be harvested from museum specimens preserved in ethanol or even dried in some cases, such as the Neanderthal genome. The difference in this case is that we have a reliable human genome, which is closely related to the Neanderthal, with which to compare it. For many plant and animal genomes of interest, there is no reference to which a new genome can be compared, making the assembly and analysis that much more challenging.

Bioinformatics Challenges

There are also bioinformatics challenges at every step, starting with assembling genomes from DNA sequence data and proceeding through annotation, comparison, and phylogenetic analysis. Traditional cluster computing environments have been a mainstay of scientific computing because they can distribute large, complex problems across many CPUs. Historically, these clusters have used hard disk drives (HDD) for storage and highly distributed RAM. Genome assembly, on the other hand, requires large amounts of shared or dedicated RAM. Genome assembly is often slow, impractical, or impossible on traditional clustered infrastructures.

To keep pace with rising data volumes and processing requirements, genomics researchers need affordable, scale-up platforms that provide greater memory capacity, higher performance, faster throughput, greater density, and outstanding energy efficiency.

"As DNA sequencing has become vastly more cost- and time-efficient, the key challenge is analyzing the huge amount of raw genomic data, which requires tremendous computing power. The Smithsonian is utilizing Intel's latest processing and storage technologies, enabling us to accelerate our ability to uncover biology from genomes of diverse species across the tree of life.

Advances in genome sequencing are leading not only to the development of precision medicine for humans, but also are giving us the opportunity at this critical moment to understand and preserve all biodiversity on Earth."

Rebecca Dikow
Smithsonian Institute for Biodiversity Genomics

At the Smithsonian: Biodiversity Across the Tree of Life

Reflecting the importance of genomic analysis to sustaining a diverse planet, the Smithsonian Institution is taking a leadership role in biodiversity genomics. Established in 1846, the Smithsonian has 19 museums which are among the most visited in the world, and its nine research centers and research projects span the globe.

Collection has long been a core element of the Smithsonian's mission. The Smithsonian's National Museum of Natural History alone has 127 million specimens and artifacts,[5] and the Smithsonian is constantly adding to its vast and highly diverse collection of genome-quality samples. Smithsonian researchers are the world's experts on organisms across the tree of life—both understanding rare and previously undescribed species, and knowing how and where to find them. Although a majority of the Smithsonian's samples were collected before today's rigorous, genome-quality storage methodologies became available, they continue to provide a foundation for our understanding of the natural world.

Now, the Smithsonian is applying its expertise to addressing the challenges of biodiversity genomics, including sample availability. The Smithsonian's Global Genome Initiative (GGI), created in 2011, focuses on preserving and studying genomic diversity and increasing access to genomic information about plants, animals, bacteria, and other species across the tree of life. With collaborators around the world, GGI researchers are collecting genome-quality samples from coral reefs, tropical rainforests, and ecosystems across the world. These samples are housed in special bioreposi-tories that will preserve the DNA and RNA for future research. Smithsonian researchers are also developing innova-tive techniques that let them sequence whole groups of organisms and entire ecosystems—from mammals down to fungi and algae.

Providing further leadership, the Smithsonian in 2014 established the Smithsonian Institute for Biodiversity Genomics to set in place the tools and technologies for studying the genomic basis of life on earth. The genome data generated by Smithsonian researchers will be made publicly available for download.

In addition, researchers at the institute are planning to make genomes available for community annotation, so that experts on particular genes or gene families can contribute to the genome annotation. They're exploring this collaborative process with the genome of the golden-collared manakin, a threatened bird that lives in tropical and subtropical forests of Colombia, Costa Rica, and Panama, and are exploring ways to scale this capability.

Exploring Next-Generation Platform Technologies for Genomics

Given the drawbacks of the traditional cluster environment, Smithsonian researchers were eager to see what recent technology advances might mean for their ability to handle their fast-growing genome assembly workloads. A team of Smithsonian researchers worked with Intel to assess the performance impact of the latest Intel technologies. They used a single Intel® processor-based server that combines three crucial components:

- **Intel Xeon processor E7-8890 v3** (One four-socket processor with 18 cores per socket, for a total of 72 physical cores). Intel's latest high-end processors offer outstanding performance, price/performance, energy-efficiency, and capacity for data-intensive workloads. The large number of very fast CPUs, coupled with fast storage, expanded memory capacity, and an innovative memory acceleration architecture, are particu-larly powerful for genome assembly and analysis. In addition to high performance and throughput, they offer the manageability and cost savings of a single, compact system to manage.
- **DDR4 RAM** (1 TB). This double-data-rate, fourth-generation dynamic memory technology enhances performance and energy efficiency for memory-intensive computing. The Intel Xeon processor E7-8890 v3 supports up to 6 TB of DDR4 memory in a four-socket configuration.
- **Intel SSDs DC P3700 drives** (Four dedicated SSDs, each with 800 GB of non-volatile memory). Intel SSDs for the PCI Express* (PCIe*) bus support the Non-Volatile Memory Express* (NVMe*) standard, an open interface developed by Intel and other vendors to enhance storage performance by taking advantage of PCIe's low latency and parallelism. Intel Data Center Family SSDs are rigorously engineered and tested to ensure reliability for enterprise-class workloads. The dedicated SSDs enhance throughput by acting as cache accelerators and reducing the need to access slower network storage.

Smithsonian researchers use standard open source tools to support their genomics pipeline (Figure 1). For the performance testing, the team moved the portions of the assembly workflow that are within the dotted lines of the figure to the new Intel platform. They performed no special tuning.

The Smithsonian chose the *Maylandia zebra* fish dataset because it had been the subject of a previous study on the best practices of genome assembly[6] and because, at 1 Gbp, it has a "medium" genome size. Starting from 542.9 GB of raw data, the assembly consumed nearly 2.5 TB of final storage space. Peak RAM usage reached 290 GB.

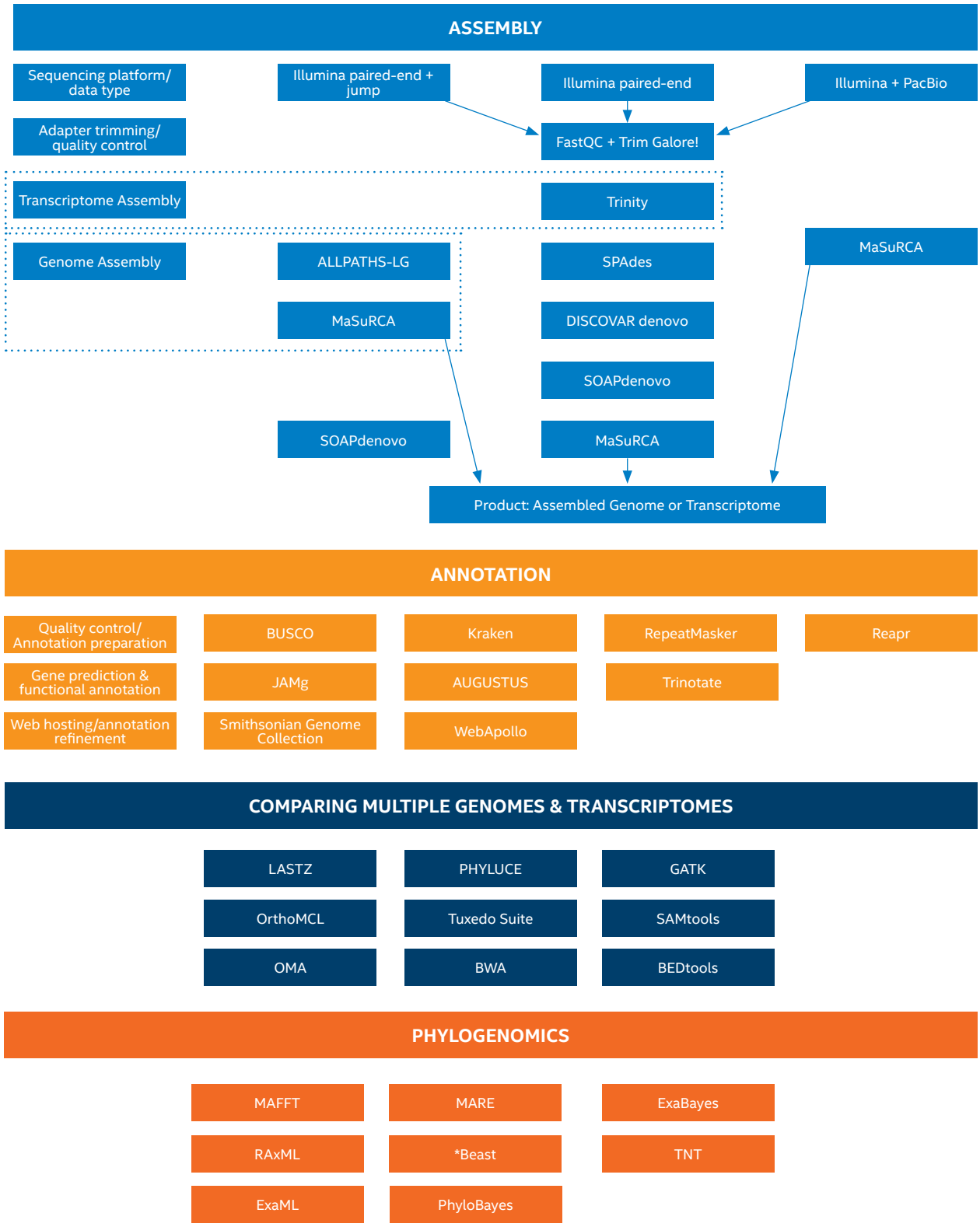

Figure 1. Genomics Workflow and Tools at the Smithsonian Institute for Biodiversity Genomics[7]
Dotted lines show work performed on the system in the initial phase of the project.

Table 1 summarizes the test results, showing that the 72-core server with the Intel Xeon processor E7-88990 v3 and Intel SSDs assembled this fish genome in 47 hours. This was nearly half the 90.86 hours needed using a subset of the Smithsonian's cluster, which mixed multiple generations of multiple vendors' processors and used conventional networked storage.

Illustrating the impact of the Intel SSDs, the assembly time on the Intel® platform with the SSDs was more than 50 percent faster than the same Intel-based server using networked storage. Reflecting the researchers' needs to process diverse workloads and genome sizes, they also tested smaller portions of the dataset and found similar speedups.

These improvements in assembly performance can help researchers at the Smithsonian Institute of Biodiversity Genomics accelerate results for genomes of varying sizes and species and for the full genomics pipeline. Along with the potential improvements in floor space, energy efficiency, and manageability offered by the Intel platform compared to the traditional cluster, these results represent an exciting opportunity for Smithsonian biodiversity researchers to reduce infrastructure costs while accelerating vital scientific progress.

Next Steps

Based on its findings to date, the research team is experimenting with simultaneous analysis of multiple data sets. We are also moving on from the fish genome to other similar-sized genomes such as a bird (Manacus vitellinus) and then to larger genomes (e.g., mammals and plants). We plan to compare diverse genomes with a goal of providing hardware recommendations for the biodiversity genomics community at large.

The researchers believe the Intel platform has the potential to enable the assembly and analysis of genomes for as many species as Smithsonian scientists can gather data. By testing the hardware on a diverse set of genomes, we can begin to understand what kinds of impacts genome biology has on genome assembly and what recommendations we can give to organizations that are undertaking their own genome sequencing projects.

Smithsonian scientists are continuing their basic research and cross-disciplinary, collaborative work to advance biodiversity genomics and apply genomics data to important challenges. This work will help scientists, researchers, policymakers, and others:

- Increase knowledge and understanding of the changes occurring on the planet
- Enable more informed decision making about climate change, species and ecosystem preservation, and other critical challenges
- Increase knowledge of evolutionary history and physiology
- Preserve dying species for possible revival in the future
- Contribute to breakthroughs in areas such as medicine, energy, agriculture

System Configuration	Small Dataset (2 Libraries 1 Paired End, 1 Mate Pair – 73.6GB)	Complete Dataset (8 Libraries - s 1 Paired End, 6 Mate Pair, 1 fosmid Pair 542.9GB)
Subset of the Smithsonian shared cluster (3,300 cores with mixed processors, 16 TB of RAM, networked storage)	7.50 Hours	90.86 Hours
Intel® Xeon® processor E7-8890 v3 (72 cores, 1 TB DDR4 RAM, networked storage)	6.15 Hours	75.0 Hours
Intel Xeon processor E7-8890 v3 (72 cores, 1 TB DDR4 RAM, Intel® SSD P3700 drives)	4.02 Hours	47.0 Hours

Table 1. Genome Assembly Speedups: Intel® Xeon® processor E7-8890 v3, Intel® SSDs+.

+ Tests conducted by the Smithsonian Institute for Biodiversity Genomics using ALLPATHS-LG build 52415 and Gnu Compiler Collection (gcc) 4.9.2. The cluster ran CentOS 6.6, and the Intel® Xeon® processor-based server ran CentOS 6.6 and Red Hat Enterprise Linux (RHEL) Server 7.1 (Maipo).

Enabling Advances

Computing resources are key enablers for future genomic breakthroughs, from steps to cure human cancers to those that will help preserve biodiversity. Intel continues to advance its platform technologies to meet the intense demands of genomics researchers. Of particular note for big data users in the life sciences, Intel and Micron have announced Intel® 3D XPoint™ (pronounced cross-point) technology, a forthcoming disruptive innovation in non-volatile memory that is 8-10 times denser than conventional DRAM memory and up to 1,000 times faster than today's NVM technologies.[8]

Intel also collaborates with the open source community, sequencing companies, and leading users such as the Smithsonian Institution, to understand life science computing requirements, advance algorithm development for the life sciences, and ensure practical, end-to-end solutions.

By building their genomics infrastructure on Intel technologies, whether in on-site infrastructure or on external clouds, life science researchers are better able to keep pace with the rising demand for informatics capacity and performance, benefit from ongoing improvements in energy efficiency and density—and advance their crucially important science.

Smithsonian scientists described more than 500 new species last year. With the speedups now possible in genome analytics, researchers envision being able to publish a species' DNA when they announce its identification.

Making a Difference with Biodiversity Genomics

With fast, affordable, high-capacity platforms for the biodiversity genomics pipeline, scientists can bring genomic data to bear on a wide range of questions and challenges. Here are a few areas that can benefit:

Disease. A fungal disease is one of several conditions wiping out amphibian populations around the world—but some species and individuals are less susceptible to the disease than others. Researchers at the Smithsonian Conservation Biology Institute and National Zoo are sequencing frog transcriptomes from multiple tissue types to study genomic influences. By understanding the differences in gene expression between resistant and susceptible amphibians, scientists will be better able to reduce die-off. They may also gain insights that will apply to other species, including humans.

Population. How did a population evolve? Which genes influence or determine important population traits? To answer those questions, scientists need to understand the population's structure, and to do that, they must sequence large numbers of genomes from within a given population. Population studies, such as those being conducted at the Smithsonian Tropical Research Institute to understand population traits of Heliconius butterflies, can contribute insights into the best ways to preserve diversity within and between populations.

Conservation. The Panamanian golden frog is among the thousands of species that face extinction. Sequencing the genomes of these endangered organisms and species before they become extinct can enhance captive breeding programs and other approaches that may prevent extinction. Genome data can also preserve knowledge of the species, answer questions such as why frog genomes are so enormous (up to 20 Gbp). It can also improve management and planning decisions, and increase understanding of the frogs' genomic diversity.

New species. Scientists from the Smithsonian National Museum of Natural History described more than 500 new species last year. Publishing the new species' genome along with the species description will put scientists everywhere much further along the path to knowledge and understanding of the species, its environment, and its potential impact on humans.

Metagenomics. How do gut bacteria influence health? How is ocean life being altered by climate change and industrial pollution? Answers to those questions may come from advances that make it possible to sequence all the genomes present in a given location—whether it's a human's gut or a sample of the seawater surrounding a tropical reef. Scientists at the Smithsonian Marine Global Earth Observatory can shorten time-to-insight by applying genomic analysis to their studies of coastal marine biodiversity and ecosystems.

Learn More

Learn more about 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

Read about the Smithsonian Institute for Biodiversity Genomics:

- http://biogenomics.si.edu

Stay abreast of continued innovations. Follow us on Twitter:

- @portlandketan, @IntelHealth
- @smithsonian

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

- https://communities.intel.com/community/itpeernetwork/healthcare

[1] Next Gen Seek: How Does a Single HiSeq X Compare with HiSeq 2500? January 14, 2014. http://nextgenseek.com/2014/01/how-does-a-single-hiseq-x-compares-with-hiseq-2500/

[2] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical? PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195. http://www.plosbiology.org/article/fetchObject.action?uri=info:doi/10.1371/journal.pbio.1002195&representation=PDF

[3] For an Intel white paper highlighting plant and animal genomics at the US Department of Agriculture, see Advances in Sequencing Drive Plant and Animal Genomics to the Cloud, Intel, 2015. http://www.intel.com/content/www/us/en/government/long-read-sequencing-drives-genomics-to-the-cloud.html

[4] Rachel Nuwer, Extinction Rates are Biased and Much Worse than You Thought, September 2, 2012. http://www.smithsonianmag.com/smart-news/extinction-rates-are-biased-and-much-worse-than-you-thought-24290026/?no-ist=

[5] Smithsonian National Museum of Natural History Factsheet. http://newsdesk.si.edu/factsheets/national-museum-natural-history

[6] Bradnam et al.: Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2013 2:10. http://www.gigasciencejournal.com/content/2/1/10

[7] See Figshare for software mentioned in the figure. Dikow, Rebecca (2015): Smithsonian Genomics Workflow. figshare. http://dx.doi.org/10.6084/m9.figshare.1588781

[8] Technology claims are based on comparisons of latency, density and write cycling metrics amongst memory technologies recorded on published specifications of in-market memory products against internal Intel specifications. Learn more about 3D XPoint technology at http://www.intel.com/content/www/us/en/architecture-and-technology/3d-xpoint-technology-animation.html and http://newsroom.intel.com/community/intel_newsroom/blog/2015/07/28/intel-and-micron-produce-breakthrough-memory-technology.

Printed in USA 0116/DW/HBD/PDF Please Recycle 333807-001 US