TECHNICAL BRIEF
www.cray.com/solutions/life-sciences

HAIL POWERS PRECISION MEDICINE ON THE URIKA®-GX PLATFORM

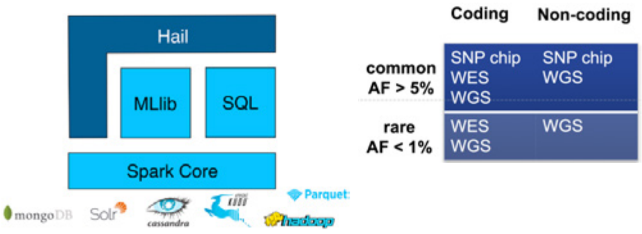
The Cray® Urika®-GX platform's large RAM capacity and best-in-class Aries™ interconnect can be transformative in the HTS ecosystem and help users realize significant performance gains.

The Power to "See"
Hail is an open-source platform for analyzing variants in genomic data on top of Apache Spark™ (see Figure 1a). It takes advantage of three key elements of Spark's design:

- **Scalability.** Datasets are multi-terabyte and growing rapidly.
- **Simpler APIs.** They hide the complexity of distributed computing and parallel execution, and let biologists explore data using familiar biological terms.
- **Algorithms for large-scale linear algebra and ML.** Performant code leverages both linear algebra legacy and custom libraries (for example, Cray compilers and tools).

Hail provides a parallel Scala API as well as parallel Python API, and adds powerful, expressive high-level layers that include: **fast, easy data ingest** in various data formats, especially if you're using parallel file I/O from file systems like Lustre®; expressive methods to manipulate and visualize high-dimensional data; and statistics and ML methods specific to apps in genetics.

Figure 1b summarizes the power of three prevalent variant analysis technologies – genotype arrays (aka SNP chips), whole exome sequencing (WES) and whole genome sequencing (WGS). Each SNP chip sample typically includes about 1 million SNPs. WES only covers coding regions – approximately 1.5 percent of whole genome. For common variants – those with allele frequencies larger than 5 percent – all three technologies are effective at detecting variants in coding regions, and WES cannot resolve noncoding mutations. For rare variants – those with



	Coding	Non-coding
common AF > 5%	SNP chip WES WGS	SNP chip WGS
rare AF < 1%	WES WGS	WGS

Figure 1a (left). Hail built atop Apache Spark. Figure 1b (right). Summary of the power of three prevalent variant analysis technologies – genotype arrays (SNP chips), whole exome sequencing (WES) and whole genome sequencing (WGS). Assuming enough samples are available, only WGS detects rare, noncoding variants.

AF less than 1 percent – SNP chips cannot resolve variants. Only WGS (currently almost 100 times bigger per sample than WES) can resolve variants in noncoding regions – if and only if a large number of samples (more than hundreds of thousands) are available. Researchers are also homing in on URVs with AF around 0.01 percent. Both challenges – rare and UR variants – are major drivers for scaling up genetic analysis with new technology to generate a detectable rare variant "signal."

API for Precision Medicine Discovery

Cray provides a set of tools to manage your data lifecycle. The cycle starts when genetic data are loaded onto fast, local storage. After variants are discovered on individual samples using GATK Best Practices workflows, Hail processes the global view of all samples at scale and outputs a summary of statistically significant variants in a VDS file (see Figure 2). Hail's variant dataset format allows independent access of genotypes on both variants and samples. Build a Hail workflow by iteratively applying algorithms and visualizing results using Jupyter Notebooks. Containerize the workflow, then deploy it using Mesos™ and Marathon. Deepen your analyses and transform your discovery science with Cray's high-performance Urika-GX platform and underlying Apache Spark engine.

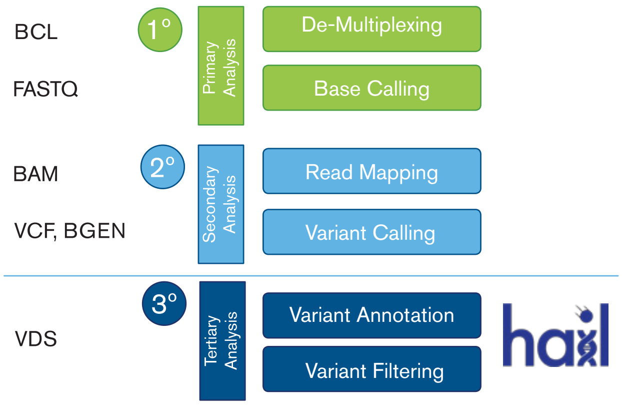


Figure 2. Genetic variant analysis workflow with Hail.

Annotating the Deluge

Genetic researchers are drowning in a deluge of sparsely annotated data – they need well-annotated variants based on as many millions of samples (one sample per human genome or exome) as budgets allow. Computational biologists use fine-grained, embarrassingly parallel ops on commodity hardware running open source software to extract variants per sample. They wrangle the hundreds of trillions of variants (aka genotypes) from hundreds of thousands of samples into fast memory to deeply understand the biology through the lens of statistics. Consortia of genetic subject matter experts (SMEs) continue to gather samples and perform quality control (QC). QC on both sample and variant spaces creates accurate annotation – labor-intensive, iterative quality control on these samples. SMEs need the ability to cycle through the entire cohort multiple times a day. Accurate annotation used in production workflows is essential for actionable, precise decisions for patients in the clinic.

Scalage

Daniel MacArthur's lab in the Medical and Population Genetics (MPG) group at the Broad Institute used Hail running on the Urika-GX platform to deliver the first public release of the Genome Aggregation Database (gnomAD) to the research community at the American Society of Human Genetics' 2016 conference. This meant scaling their capacity up fourfold, drastically reducing latency to iterate on important input subpopulations many times per day, and scaling productivity more than tenfold – transformative change for data science. Roughly a quarter of the raw input WES and WGS sequence data are removed by QC using filtering on both variants and samples.

The gnomAD team annotated the variants from hundreds of thousands of WES and WGS raw input samples for an initial release in one week. All the time-critical, low-latency processing and QC analyses for the WES varianceswere performed using Hail and other machine learning techniques on the Urika-GX platform's flexible, scalable framework

– one that proved crucial for processing such large datasets in a reasonable amount of time, allowing for exploration of the data at a much more rapid scale.

Impact

Researchers at the Broad Institute and their collaborators leveraged Hail, gnomAD and other population datasets and the power of Cray's Urika-GX platform to hugely impact genomic variant research in the last year. Nearly 30 papers and research projects using Hail across many disease types and institutes are underway. High-level support and reliability in Spark was essential for this success. The Urika-GX platform delivered for the Broad, especially when gnomAD had to be QC'd under tight deadlines. Key in this progress was the ability to fit a large dataset into memory and see where the code failed – not just faster, but *at all* – especially in the early stages of our work together.

Productivity

Filtering massive genomic datasets is hugely labor intensive and iterative. QC involves sweeping a large space of filtering methods and metrics to select subsets. Neither sample nor variant QC processing is "pushbutton" – both require subject-matter and coding expertise. Cray's Urika-GX platform exceeded the low-latency QC demands for gnomAD and vastly reduced iteration wait times. This was transformative. The net effect was sharper focus and higher productivity across many research programs – not just higher-quality code, but more effective teams of SMEs, software developers, users and managers working together.

The Urika-GX supercomputer's large RAM capacity and best-in-class Aries™ interconnect can be transformative in your NGS ecosystem and help you realize significant performance gains.

We immediately saw speedups of more than 4x in mid-2016 during our work with early versions of the GATK4 engine with Spark support for computing BQSR on WGS BAMs. An immediate 5x speedup – without tuning Hail code – was realized as Daniel MacArthur's MPG team at the Broad migrated from a small Hadoop cluster to the Urika-GX platform and commenced gnomAD QC. In less than a day, you can annotate over 140,000 WES on this platform. You can download and read more details in a Cray solution brief, "Pinpointing Mutations for Precision Medicine."

SOLUTION

Cray® Urika®-GX agile analytics platform

The Urika-GX platform fuses supercomputing with an open, enterprise framework, giving it the speed and agility to handle a variety of workloads. Among its abilities, the Urika-GX system runs Hadoop®, Spark™, graph and HPC analytics workloads concurrently.

SYSTEM DETAILS

Nodes: 32

Processor cores: 1,024

Memory: 8 TB

SSD: 22 TB

Local Disk: 128 TB

Network: Aries™ interconnect

Cray Inc.
901 Fifth Avenue, Suite 1000
Seattle, WA 98164
Tel: 206.701.2000
Fax: 206.701.2500
WWW.CRAY.COM