BIOTEAM
Enabling Science

SLIPSTREAM GALAXY EDITION APPLIANCE

SLIPSTREAM APPLIANCE: GALAXY EDITION

The Galaxy Project and BioTeam Inc. form a Strategic Alliance to Deliver a Plug-and-Play Appliance to Streamline Biomedical Research

For more than a decade, biomedical research has been steadily transformed into a data intensive science. Biomedical instruments now generate massive datasets that can only be analyzed using sophisticated computational infrastructure and often difficult-to-use informatics. The limited availability of compute resources and technical skills among life science researchers has impeded progress in both the lab and the clinic.

Next-generation sequencing (NGS) is perhaps the best illustration of this problem, with instruments generating upwards of 600 gigabases of data per run that need to be analyzed and interpreted. The recent rise of affordable NGS desktop instruments, which are expected to reach 80% of the NGS instrument market by 2015, is democratizing research but at the same time exacerbating the analysis challenges[1]. With the increasing amount of data from these desktop instruments (Figure 1) and their proliferation from large sequencing centers into smaller labs, the scarcity of informatics and IT infrastructure resources are becoming even more of a problem.

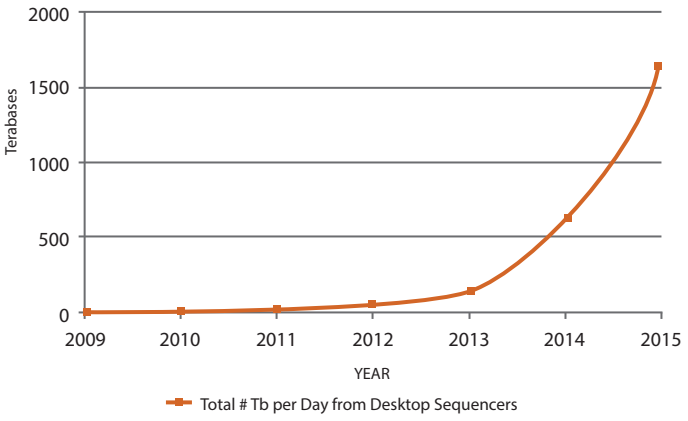

Figure 1: Growth of Data Generated by Desktop Sequencing Instruments

THE GALAXY PROJECT: SIMPLIFYING ANALYSIS

From its inception, the Galaxy Project[2] has focused on creating an open scientific analysis platform to make it easier for researchers to analyze data, especially those with less IT and informatics expertise. The Galaxy Project organizers starkly described the situation in a 2010 paper:

"[The] sudden reliance on computation has created an 'informatics crisis' for life science researchers…Without programming or informatics expertise, scientists needing to use computational approaches are impeded by problems ranging from tool installation; to determining which parameter values to use; to efficiently combining multiple tools together in an analysis chain."[3]

Though the Galaxy analysis platform was initially developed for genomics research, "it is largely domain agnostic."[4] Since its start in 2006, the Galaxy community has grown impressively. Now, with over 31,000 registered users on the Galaxy Main Server, the number of jobs run each month consistently exceeds 100,000. In 2012 alone there were over 400[5] papers citing Galaxy. By any measure, Galaxy's traction is substantial. Many prominent institutions worldwide, such as Cold Spring Harbor Laboratory, MIT, University of Oslo, and National Cancer Institute are working with Galaxy.

SLIPSTREAM GALAXY: DEMOCRATIZING ANALYSIS

The Galaxy Project and BioTeam, a highly respected group of consultants specializing in implementing informatics, IT, and high-performance computing solutions for academic and industrial life sciences organizations, have formed a strategic alliance to build and offer a preconfigured appliance - the SlipStream Appliance: Galaxy Edition (SlipStream Galaxy), seen in Figure 2. This convenient, affordable appliance is carefully designed to handle the intensive demands of biomedical data analysis and lowers the barrier for entry into data analysis for IT-constrained researchers and laboratories. (see Table 2 for appliance specifications).

Figure 2: SlipStream Appliance

Anushka Brownley, Product Manager for SlipStream at Bioteam, said that "SlipStream Galaxy is designed to reduce the IT and administrative burden of running a production instance of the Galaxy analysis platform and the underlying computational tools. SlipStream Galaxy integrates a powerful computational infrastructure and storage system in a desktop server to provide a dedicated resource to quickly analyze data with Galaxy."

BioTeam has worked with many major organizations over the past decade including NIH, the Broad Institute, and Pfizer. The team is equally at home helping smaller groups with modest resources integrate and handle the wealth of genomics data. The development of SlipStream Galaxy is a logical progression in serving the community.

> "BioTeam has years of experience doing consulting, and is one of the most well respected IT and bioinformatics consulting groups in the community. They bring not only the technical expertise with respect to building systems and software stacks for bioinformatics, but also the bioinformatics expertise that is one of the most important parts of supporting a Galaxy appliance."
> – Anton Nekrutenko, Galaxy Project Co-Founder, Penn State University

Currently, users access Galaxy in one of three main ways:

1) Via an account on the free public Galaxy server hosted at Penn State University (Galaxy Main) or a Galaxy server hosted at another institution
2) Via a cloud-based instance on Amazon Web Services
3) With local installations of the Galaxy platform

Each method has distinct benefits and drawbacks (Figure 3).

	NO WAIT TIMES	NO STORAGE QUOTAS	NO JOB SUBMISSION LIMITS	NO DATA TRANSFER BOTTLENECKS	NO IT EXPERIENCE REQUIRED	NO REQUIRED INFRASTRUCTURE
GALAXY MAIN	✗	✗	✗	✗	✓	✓
LOCAL GALAXY	?	?	?	✓	✗	✗
CLOUD GALAXY	✓	✓	✓	✗	✗	✓
SLIPSTREAM GALAXY	✓	✓	✓	✓	✓	✓

Figure 3: A comparison of the different ways to use Galaxy

Galaxy Main is the most popular method for using Galaxy. However, due to overwhelming demand, quotas were implemented. As the use of Galaxy Main has grown, so have job execution and wait times.

> "Because the amount of data being generated by sequencing technologies is so vast, a centralized service like Galaxy Main can no longer meet the community's needs. Analysis needs to be decentralized, and one great way to do this is by having local Galaxy instances so that data can be analyzed near where it is produced on dedicated hardware. Unfortunately, this presents barriers for many groups – determining and acquiring the right hardware, installing Galaxy and all of the associated tools and data, et cetera. A pre-built appliance eliminates all of these barriers, allowing research groups to run their own local Galaxy with minimal effort."
> – James Taylor, Galaxy Project Co-Founder, Emory University.

SlipStream Galaxy is a natural complement to desktop NGS instruments, and also serves the broader purpose of providing a self-contained, easy-to-use IT appliance for any biomedical data analysis application. It is pre-loaded with a fully functional Galaxy instance and is, therefore, able to accommodate a wide range of research.

THE (DO NOT HAVE) TO-DO LIST

SlipStream Galaxy seamlessly bridges the divide between researchers without informatics expertise and the complex computational infrastructure they need to perform their work. This point is made clearly by the Galaxy Project team:

"The most important feature of Galaxy's analysis workspace is what users do not need to do or learn. Galaxy users do not need to program nor do they need to learn the implementation details of any single tool. Galaxy enables users to perform integrative genomic analyses by providing a unified, web-based interface for obtaining genomic data and applying computational tools to analyze the data. Users can import datasets into their workspaces from many established data warehouses or upload their own datasets. Interfaces to computational tools are automatically generated from abstract descriptions to ensure a consistent look and feel." [6]

This same philosophy characterizes SlipStream Galaxy. By adopting this appliance, IT-constrained researchers are relieved of most underlying IT-infrastructure tasks such as hardware evaluation, purchase, set-up, and operating system and database installations. BioTeam estimates that by using SlipStream Galaxy to get up and running, the average lab can save up to one month of deployment time with a start-up cost savings (typically charge-backs to internal IT staff) that is easily over $20,000.

Here are a few key IT tasks SlipStream Galaxy users do ***not*** need to undertake to deploy and manage a dedicated Galaxy server:

Hardware Specification
The appliance has already been specified – CPUs, RAM, network capability, storage, and more – to best handle large NGS datasets and is configured to accelerate analysis computation.

OS Configuration
Tricky OS installation and file system configuration are completed and tested.

Database Installation and Configuration
Installation of database with replication and advanced proxy configuration are done.

Galaxy Job Management
Installation and configuration of both job management software and Galaxy, and implementation appropriate resource management are complete.

Galaxy and Analysis Tool Installation
Latest, stable versions of Galaxy and all of the underlying computational tools accessible through the default Galaxy interface are pre-loaded on the appliance.

Automated Updates
Galaxy and all of the underlying computational tools can be automatically updated as new versions are made available.

Dataset Downloads
Commonly used reference datasets for use with Galaxy, such as human, mouse, the *D. melanogaster, C. elegans,* and *S. cerevisiae*, are pre-loaded. Additional datasets may be made available upon request.

Performance Tuning
The Galaxy software and SlipStream hardware have been tuned for faster uploads and downloads.

DELIVERING PERFORMANCE

Feedback from early development partners has shown the SlipStream appliance to be a highly capable analysis platform. "Sequencing is so much easier now, approaching trivial really. A device that centralizes functions with respect to data archival, storage, and analysis is a tremendous aid," says Ed DeLong, a prominent metagenomics researcher at MIT and member of the National Academy of Sciences. His lab has used the SlipStream appliance for several months.

> "Something that surprised us was that the amount of memory built into the device and its considerable compute capability have allowed us to run substantial jobs directly on the device which previously we would have had to run on an MIT 500 core cluster. Because it's all in one device it's much easier to manage. Typically if you are trying to run a compute cluster and you have other storage devices and your sequencer, you are running multiple different operating systems which entails a lot of systems administration and IT management overhead."
> – Ed DeLong, PI, MIT

TOOLS	TASK	DATA	RUN-TIME
Bowtie 2	Mapping whole human genome	204 million paired-end 100bp Illumina reads	2 Hours 44 Minutes
SAMTools	SAM-BAM conversion	127GB SAM (41GB resulting BAM)	2 Hours 7 Minutes
TopHat 2	RNA-Seq mapping	24 million 100bp Illumina reads	1 Hours 24 Minutes
Cufflinks 2	Differential Expression Analysis	4.3 GB SAM File	0 Hours 11 Minutes

Table 1: Performance Benchmarks for SlipStream Galaxy

The SlipStream hardware is specifically designed to reduce the cost and IT support requirements and optimized to accelerate NGS analysis. The benchmarks (Table 1) that were generated demonstrate the high level of performance that can be achieved by SlipStream Galaxy.

With hundreds of tools currently in the Galaxy tool shed[7], among them numerous NGS tools and workflows for short-read mapping, ChIP-seq, RNA-seq, metagenomics, and variant analysis, there are few limits to what can be done. The powerful SlipStream hardware has been configured to finish a typical computationally intensive analysis job in under twelve hours, which is convenient for researchers who want to run jobs overnight.

OPEN SOURCE, OPEN PLATFORM

BioTeam is a longtime supporter of and contributor to the open source community. BioTeam founder Chris Dagdigian serves as the main administrator and a board member of the not-for-profit Open Bioinformatics Foundation, which supports projects including BioPerl, BioJava, BioRuby and BioPython. BioTeam has been especially active in the Grid Engine project, contributing to the code base, hosting and maintaining the GridEngine information site, and offering technical support.

Along those lines, SlipStream Galaxy is designed to be an open platform. Users will have administrative access to use the system as they see fit, beyond using Galaxy. Users can install and run new applications, develop new software, and store and manage custom data. In addition to that, a portion of the profits from SlipStream Galaxy will be donated to the Galaxy Project to support ongoing development and growth of the community.

A GATEWAY TO ADDED RESOURCES

On top of being open, the SlipStream appliance is designed to seamlessly interoperate with additional computational and storage resources. As a result, the appliance can easily leverage alternate applications and infrastructure. The following are some examples of the added functionality that can be enabled.

Leverage Local Infrastructure

Scale the SlipStream appliance's compute and storage resources by bursting to your local HPC and storage infrastructure.

Expand Storage

Add storage capacity directly to the SlipStream appliance to handle increases in data volume.

Connect to the Cloud

Scale storage resources and build archives in the cloud or burst into the cloud for additional compute resources.

Software Integration

The open platform enables any type of software to be installed, integrated or to interact with the SlipStream appliance.

Federate Multiple Appliances

Leverage many SlipStream appliances as a collective resource to facilitate shared compute, data, applications.

With these capabilities, the SlipStream appliance can become much than a powerful stand-alone system. By integrating necessary resources with the SlipStream appliance, it can serve as a platform for consolidation and provide a single entry point, or gateway, to those resources.

INTEL AND BIOTEAM COLLABORATION

BioTeam is collaborating with the worldwide technology supplier Intel Corporation on a broader strategy for the SlipStream Appliance product line beyond SlipStream Galaxy.

"We have been working with BioTeam for about a year on this project," says Ketan Paranjape, Intel Global Director, Healthcare and Life Sciences. "One of Intel's core strengths is expertise optimizing hardware and software for data-intensive computation. We're able to maximize the benefit gained from using Intel Architecture such as ensuring features on the Xeon processor are used optimally for a particular task. Intel can do this at all architecture levels (CPU, storage, networking, full systems, OS and compilers, etc.). The SlipStream Appliance was architected to deliver power, expandability, and affordability, all critical must-haves in life sciences and healthcare."

PARTNERING WITH THOUGHT LEADERS

BioTeam is introducing a limited availability early access program strategically designed for organizations willing to partner with BioTeam to continue to enhance SlipStream Galaxy. BioTeam will work closely with these few selected partners and provide them with significant added value above and beyond that of the SlipStream Galaxy itself.

Seamless Adoption.
BioTeam will work closely with partners to integrate SlipStream into their environment.

Dedicated Support.
BioTeam will provide customized technical input on infrastructure and data analysis to optimize analysis using Galaxy.

Workflow Generation.
BioTeam will work with partners to manage the design of workflows and their integration into the public domain.

Optimized Utility.
Partners will have the opportunity to contribute to the utility, design, and implementation of the appliance, a key component to the Galaxy Project's offering.

SlipStream Galaxy, with its carefully designed high-performance hardware and pre-configured Galaxy software, is being offered for the extremely affordable price of **$19,995**. As it becomes more difficult to secure funding to support biomedical research, it is natural to evaluate whether the cost of such an appliance can be justified. Tristan Lubinski, a graduate student at Boston University, was an early tester of SlipStream Galaxy and used the system to analyze data for his lab in addition to generating the benchmarks in Table 1.

> "People will likely start out thinking, why would I spend money when these are free analysis tools and I have access to Galaxy Main, which is also free. But after having spent time moving files, waiting in queues for jobs to run, and dealing with old software versions, it becomes very evident how nice it is to have one of these to yourself."
> – Tristan Lubinksi, Graduate Student, Boston University

The early access program will continue until the general release of SlipStream Galaxy, which is currently planned for late 2013.

RELYING ON WORLD-CLASS SUPPORT

One missing element not specifically available from the Galaxy Project or the Galaxy community is extensive on-demand support. This is a role that BioTeam will fill. Based on feedback from the early access program, BioTeam will develop a commercial-grade support offering to be made available with the general release of SlipStream Galaxy. According to James Taylor, "While Galaxy enables researchers to run analysis easily, one of the biggest challenges they still face is determining what the right way to analyze their data is. We as an open source project are able to provide some support, but only through mailing lists and without any guaranteed turn-around. BioTeam's expertise in bioinformatics analysis will bring a level of support, for those who want it, beyond anything offered before."

LOOKING AHEAD

In practice, SlipStream Galaxy will fulfill multiple roles. For many labs it will be the primary biomedical research analysis platform. Others may use it as a development sandbox to prepare workflows for later use on larger clusters or in the cloud.

Data-intensive biomedical research is no longer new, yet the IT and bioinformatics challenges associated with making sense of the massive data sets persist. In one recent study[8], improved usability of bioinformatics tools was the top unmet need for both high throughput and low throughput instruments. SlipStream Galaxy provides a formidable IT infrastructure and analysis platform to drive forward data intensive fields, including NGS.

For more information about the SlipStream Galaxy and the early access program, contact:

slipsteam-galaxy@bioteam.net or visit
www.bioteam.net/slipstream/galaxy-edition

[1] DecBio NGS market report, http://www.decibio.com/NGS

[2] The Galaxy Project, http://galaxyproject.org;

[3] *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences*, Goecks et al. Genome Biology 2010, 11:R86 (http://genomebiology.com/2010/11/8/R86)

[4] http://en.wikipedia.org/wiki/Galaxy_(computational_biology)

[5] Galaxy Main Usage, http://wiki.galaxyproject.org/Galaxy%20Project/Statistics#Membership

[6] *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences*, Goecks et al. Genome Biology 2010, 11:R86 (http://genomebiology.com/2010/11/8/R86)

[7] http://toolshed.g2.bx.psu.edu

[8] DecBio NGS market report, http://www.decibio.com/NGS

SlipStream Appliance: Galaxy Edition Product Specifications

SOFTWARE AND TOOLS	
Galaxy	Production configuration of Galaxy
Galaxy Tools	Assembly, annotation, variant calling, and other analysis tools accessible through the default Galaxy interface
Galaxy Datasets	Human, mouse, the *D. Melanogaster, C. elegans,* and *S. cerevisiae*. Additional datasets may be available upon request
HARDWARE	
CPU	2x Intel® Xeon® Processor E5-2690, 8 core (16 cores total)
Memory	12x 32GB RDIMM (384GB) Optional Upgrade to 512GB
Storage	7x 3TB SAS 6 Gbps HDD (16 TB usable) 1x 100GB Solid State Disk
Power	Dual Redundant Power Supplies
Network	Dual Gigabit Network Adaptor
SUPPORT	
Installation	Includes setup
Warranty	Includes 36 month hardware warranty
Annual Support (Additional, available after general release)	Galaxy and analysis tools upgrades + basic support Premium Galaxy support
Customization (Additional)	Custom services such as interface customization, process automation, system integration, and more

Table 2: SlipStream Galaxy Product Specifications