



Broad Institute Develops GenomeSpace Platform to Integrate -Omics Tools, Enable Reproducible Workflows

July 06, 2009

Newsletter: [BioInform](#)
[BioInform - July 6, 2009](#)

By [Vivien Marx](#)

This article has been updated from a previous version to clarify the development history of GenePattern.

STOCKHOLM, Sweden — The Broad Institute is following on its GenePattern analysis suite with a platform called GenomeSpace that will integrate six open source bioinformatics tools and eventually enable other developers to add their own software.

Jill Mesirov, associate director and CIO at the Broad, discussed the project here this weekend at a satellite meeting held in conjunction with the joint Intelligent Systems for Molecular Biology and European Conference on Computational Biology conferences.

Mesirov called GenomeSpace the "next logical step" after GenePattern — a suite of genomic analysis tools for gene expression analysis, proteomics, SNP analysis, and other bioinformatics applications that the Broad initially launched in 2004. A key aim of the project is to help researchers build and share reproducible bioinformatics workflows.

GenomeSpace will obviate the need for manual data transfer between tools and will track "all the provenance and history" of the data and the analysis performed, Mesirov said. Scientists will be able to replay their workflow at a later date or share it with others.

The plan is to initially integrate six tools: the UCSC Genome Browser, the data management and analysis tool Genomica, the pathway visualization tool Cytoscape, the Galaxy data-integration platform, and the Broad's own Integrative Genomics Viewer and GenePattern.

Mesirov said the Broad developers chose these tools for the initial implementation because of their wide use and the wide range "of architectural models" they represent, but noted that the plan is to eventually enable other developers to easily add their tools to GenomeSpace to create a "web 2.0" community for bioinformatics.

Down the line, Mesirov said, other tools, including commercial software, might be integrated with GenomeSpace, but for now the team has its hands full integrating this tool set.

The different architectures present a challenge, she said. For example, Galaxy and the UCSC Browser are web applications, while GenePattern is a client-server application, and Genomica is a closed Java application.

When the creators of these tools were asked if they wanted to participate in GenomeSpace, Mesirov said the universal answer was, "You bet." The idea is not only to make this a community project but also keep the "barrier to entry as low as possible," she said.

The programs will be integrated with the intent of retaining their "look and feel" for users familiar with them so they "don't have to learn a whole new interface."

The concept involves a way to offer "self-service" of computational tools and to create workflows in a way that does not require intervention by an informatician and allows for reproducibility, she said.

She noted, however, that GenomeSpace will also offer an entry point for programmers as well, since "we like to micro-manage."

GenomeSpace seems to have arrived at the right time for a number of tools developers. As the Penn State University team behind Galaxy states on its web site, "The methods section of too many papers sound like the data were analyzed using a collection of in-house scripts."

"Any attempt to allow users to have another way to analyze data without worrying too much about formats and compatibilities is a good thing, and I think that is the idea behind GenomeSpace," Anton Nekrutenko, associate professor of biochemistry and molecular biology at Penn State University and Galaxy co-PI told *BioInform*.

This project is about being compatible with tools that were not part of the original Galaxy, such as Cytoscape, and systems biology tools, he said.

While scientists have an increasing number of tools at their disposal, the challenge is that there is a growing gap in the use of many tools and the difficulty to get them to "reason together," Mesirov said.

GenePattern was the Broad's first response to this idea. GenePattern 3.2, launched last week, now has more than 120 modules. The system is built on a web services model that offers analytical tools alongside record-keeping features such as executable records of all sessions and the ability to save and share pipelines.

The Broad is currently collaborating with Microsoft to create a plug-in that will allow scientists to add a GenePattern workflow or pipeline, in an executable format, into Word documents in order to embed reproducible workflows in published papers. The plug-in should be available some time next month, Mesirov said.

'Empowering' Biologists

The GenomeSpace project will first focus on interoperability. The interface will allow users to navigate tools and tag them as "favorites," and a "common layer for interoperability," or CLIO, bar will let users move data back and forth. "That's the big engineering job," Mesirov said. "The onus is on us to make [the tools] speak to each other."

Within the platform, the tools will have "close relationships" so data can be sent back and forth.

GenomeSpace grew out of the reality that many smaller labs are facing, Mesirov said. While the Broad has a large computational staff, researchers at many other labs do not have on-staff bioinformaticians at their disposal and need to handle analysis themselves. GenomeSpace should help "empower" these biologists and help them think about their project differently without the need for a bioinformatician, she said.

More and more experimental biologists are using computational tools, said Mesirov. "Our job is to make that as easy as possible."

In Mesirov's view, the driving concept behind the project is reproducibility.

She explained that a big impetus came from the Broad Institute's experience 10 years ago following the publication of the Golub *et al* [Science paper](#) that was one of the first studies to use microarray data to differentiate different forms of cancer.

Following this work, Mesirov said the authors received "hundreds of e-mails" from colleagues who reported that they could not reproduce the results.

Part of the reason, she said, had to do with the study's analytical method, which was a multi-step workflow that included file preparation, processing, filtering, statistical analysis, modeling, cross-validation, and verification against separate data sets. Building the model required "stringing together" all the tools and then importing the data into Excel to visualize the results. It was a "tremendous" amount of work, with many manual steps of data transfer, she said.

The only way to capture the parameters of the workflow was through "our memories and some handwritten notes," Mesirov said. That limited the application options of the paper, even though the scientists had documented the method.

GenePattern, which encapsulated the workflow used for the Golub *et al*. study, was the Broad's first response to this challenge. Mesirov said that GenomeSpace will build on this concept by improving the interoperability of these tools.

David Rocke, biostatistician and the UC David School of Medicine, told *BioInform* that reproducibility and transparency in data management and analysis "is an ultimate necessity" for bioinformatics.

Genomeweb system

These settings are generally managed by the web site so you rarely need to consider them.

Issue Order: 1

