The GenePattern Notebook Environment

Michael Reich[1], Thorin Tabor[2], Ted Liefeld[1], Barbara Hill[2], Helga Thorvaldsdottir[2], Jill P. Mesirov[1]

[1] University of California, San Diego, La Jolla CA 92093 Email: mmreich@ucsd.edu
[2] Broad Institute, Cambridge MA 02142

Interactive notebook systems have made significant strides toward realizing the potential of reproducible research, providing environments where users can interleave descriptive text, mathematics, images, and executable code into a "live" sharable, publishable "research narrative." However, many of these systems require knowledge of Python, R, or another programming language and are therefore out of the reach of non-programming genomic researchers.

To address this gap, we have developed the GenePattern Notebook environment, which connects the popular Jupyter Notebook system to the GenePattern platform for integrative genomics, making the hundreds of bioinformatics analysis methods in GenePattern available within the powerful and flexible Jupyter Notebook interface. GenePattern provides domain-specific methods for the analysis of gene expression (RNA-seq and microarray), sequence variation, proteomics, and genomic networks, as well as general machine learning methods for clustering, classification, and dimensionality reduction, and utility methods for data import, formatting and preprocessing. GenePattern includes methods adapted from many popular open source libraries, including Bioconductor and scikit-learn. Through its module-based architecture, GenePattern also allows users to easily add their own tools, automatically making them available within the notebook environment.

The GenePattern Notebook environment consists of a plugin that provides access to GenePattern from within Jupyter Notebook, a collection of notebook documents that demonstrate the use of GenePattern Notebook in various analysis scenarios and which can be adapted for a researcher's individual use, and a web site (in development) where users can create, share, run, and publish their own notebooks. The plugin is implemented as an extension to the Jupyter Notebook system that provides a new "GenePattern" cell type, which allows users to log in to any available GenePattern server, select any available analysis, and configure and execute it. Results are available to downstream analyses, which can be either additional GenePattern method cells or standard Jupyter Notebook code cells. GenePattern Notebook is currently available as a Docker image and as an installation from the Python Package Index (PyPI).

This talk will describe how the GenePattern Notebook environment extends the analytical and reproducible research capabilities of Jupyter Notebook and GenePattern, how using GenePattern Notebook obviates the need for coding in many analysis cases, and will demonstrate genomic analyses within GenePattern notebooks.