Title
The GenePattern Notebook Environment

Permalink
https://escholarship.org/uc/item/7j12m8z4

Journal
Cell Systems, 5(2)

ISSN
2405-4712

Authors
Reich, Michael
Tabor, Thorin
Liefeld, Ted
et al.

Publication Date
2017-08-01

DOI
10.1016/j.cels.2017.07.003

Peer reviewed

Published in final edited form as:
Cell Syst. 2017 August 23; 5(2): 149–151.e1. doi:10.1016/j.cels.2017.07.003.

The GenePattern Notebook Environment

Michael Reich[1,4], **Thorin Tabor**[1], **Ted Liefeld**[1], **Helga Thorvaldsdóttir**[2], **Barbara Hill**[2], **Pablo Tamayo**[1,3], and **Jill P. Mesirov**[1,2,3]

[1]School of Medicine, University of California, San Diego, La Jolla, CA, USA

[2]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[3]Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA

Abstract

Interactive analysis notebook environments promise to streamline genomics research through interleaving text, multimedia, and executable code into unified, sharable, reproducible “research narratives.” However, current notebook systems require programming knowledge, limiting their wider adoption by the research community. We have developed the GenePattern Notebook environment, www.genepattern-notebook.org, to our knowledge the first system to integrate the dynamic capabilities of notebook systems with an investigator-focused, easy-to-use interface that provides access to hundreds of genomic tools without the need to write code.

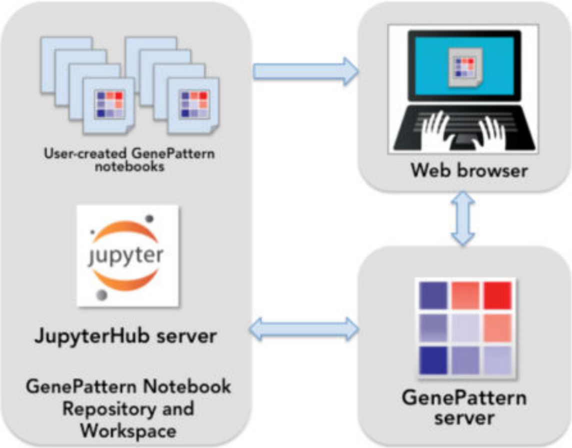
eTOC blurb

Reich et. al have developed software that integrates the capabilities of electronic analysis notebooks and bioinformatics analysis portals. GenePattern Notebook uses the popular Jupyter Notebook platform that interleaves text, graphics, and code, and brings these tools for reproducible research, as well as access to hundreds of bioinformatics analyses, to non-programmers.

Correspondence should be addressed to M.R. (mmreich@cloud.ucsd.edu).
[4]Lead contact

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Contributions
Conceptualization, M.R., T.T, P.T., J.P.M.; Software, T.T., T.L., Writing – Original Draft, M.R., T.T., T.L., H.T., J.P.M.; Writing – Review & Editing, M.R., J.P.M.; Validation, B.H.; Project Administration, M.R., H.T.; Funding Acquisition, J.P.M.


User-created GenePattern notebooks
jupyter
JupyterHub server
GenePattern Notebook Repository and Workspace
Web browser
GenePattern server

The ongoing explosion of “omics” datasets and the promise of scientific discovery arising from their analysis have given rise to software systems that aim to provide easy access to advanced methods for non-programming scientists. These “bioinformatics tool aggregation portals”, e.g., Galaxy (Afgan 2016), GenePattern (Reich 2006), and KNIME (Berthold et al., 2009), also provide for the creation and encapsulation of analytic workflows, transparent access to scalable compute resources, and removal of software installation and implementation concerns from the scientific user.

Alternatively, analysis notebook environments, inspired by the “literate programming” philosophy (Knuth 1984), integrate the exposition of a scientific project with the associated code. They aim to create an “executable document” that ideally serves as a complete description of a research project and which could also be run to reproduce the author’s results. Examples include SWEAVE (Leisch, 2002), Jupyter Notebook (Ragan-Kelley et al., 2014), Beaker (beakernotebook.com), and Zeppelin (zeppelin.apache.org).

Each of these two types of system brings significant value to its targeted user base yet has limitations that prevent wider adoption. Notebook environments model their interface around the annotation of sections of code and therefore assume that the user is fluent in a programming language such as Python or R. Bioinformatics tool aggregation portals successfully remove the requirement for coding expertise but to date have had limited ability to incorporate the variety of rich text and media formats required to represent the full scientific narrative surrounding each analysis step.

We have developed GenePattern Notebook (Figure 1), an environment that integrates the capabilities of both types of system, allowing users to incorporate encapsulated analysis tools, complete with their user-friendly interface, from a bioinformatics aggregation portal into an interactive analysis notebook. The environment is based on two long-standing software projects, the GenePattern platform for integrative genomics and the Jupyter Notebook environment for interactive computing.

GenePattern (www.genepattern.org), first released in 2004, consists of a repository of hundreds of bioinformatics analysis and visualization methods (“modules”), as well as utilities for data formatting, preprocessing, and other auxiliary functions that provide important "glue" between analysis steps. The user interface is point and click with no programming required. The public GenePattern server, hosted at www.genepattern.org since 2008, has over 40,000 registered users and runs 2000–5000 analysis jobs per week. Additional public servers are available at Indiana University (gp.indiana.edu/gp) and the Garvan Institute (pwbc.garvan.org.au/gp). The software has also been downloaded for local installation by over 17,000 bioinformatics core facilities, research laboratories, and individual scientists.

The Jupyter Notebook environment (www.jupyter.org) provides a laboratory notebook metaphor in which researchers build a step-by-step scientific narrative out of "cells" that interleaves code, formatted text, mathematical formulae, plots, and multimedia. The resulting notebooks can be shared, edited, executed, and published as complete encapsulations of *in silico* research.

The GenePattern Notebook functionality takes the Jupyter Notebook interface one step further, adding analysis, login, and rich text input components that present the GenePattern interface to provide code-free analysis and visualization (Figure S1). All cell types interact seamlessly with existing Jupyter cell types. Within a Python code cell, programming users can easily reference analysis results from a previous GenePattern analysis cell, and in a GenePattern analysis cell, programmers can use Python variables as inputs.

We integrated GenePattern with Jupyter through the use of Jupyter's *ipywidgets* package, which provides a framework for the creation of new user interface objects within Jupyter Notebooks, and GenePattern's Web services interface, which exposes all of the functionality of GenePattern (e.g., searching for and obtaining module information or querying for the execution status of an analysis) to programmatic access. This combination is a design pattern that has general applicability to the class of Web service-based tools, and the Jupyter development team is incorporating our approach into the currently evolving design of the Jupyter interfaces for graphical input. (2016, Dr. Fernando Perez, pers. comm. 26 September).

To promote the development and dissemination of GenePattern Notebooks with minimal installation requirements, we have released an online GenePattern Notebook repository and workspace where researchers can collaboratively develop and publish notebook documents. It provides a complete Jupyter environment, connections to several GenePattern servers, and for programmers, the common Python packages used in bioinformatics analysis (numpy, pandas, matplotlib, scikit, etc.). We seeded the repository with notebooks that provide commonly-used machine learning methods: clustering, classification, and prediction, as well as dimension reduction and differential expression analysis.

Those who wish to run the GenePattern Notebook environment on their own compute resources have two options: (1) Non-programmers can install the Kitematic Docker (kitematic.com) application and use it to run the GenePattern Notebook Docker image, available on the standard Docker Hub repository (hub.docker.com). This will provide a complete, ready-to-run notebook environment with all dependencies preinstalled. We also provide a Docker image for users who wish to host their own repository of GenePattern Notebooks. (2) Programmers may install the GenePattern Notebook and its dependencies through the *pip* or *conda* package manager interfaces.

To our knowledge GenePattern Notebook is the first integration of a bioinformatics tool aggregation portal with an analysis notebook environment. This approach benefits both nonprogramming and programming investigators alike. For the nonprogrammer, GenePattern Notebook provides the user-friendly GenePattern genomic analysis capabilities within a publishable notebook format. For the programmer already using the Jupyter environment, it affords easy access to the entire GenePattern library of analysis and visualization modules that can be supplemented with the investigator's own coded routines.

The GenePattern Notebook environment, along with an introductory demonstration video, documentation, and tutorials, is available at www.genepattern-notebook.org. The software is freely available under a BSD-style open source license.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, mmreich@cloud.ucsd.edu.

DATA AND SOFTWARE AVAILABILITY

GenePattern Notebook web site and online repository: http://www.genepattern-notebook.org

ADDITIONAL RESOURCES

GenePattern web site: http://www.genepattern.org

Jupvter Notebook environment: www.iupvter.org

Kitematic web site: https://kitematic.com

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by NIH grants R01-GM074024 and U24-CA194107. We thank Fernando Perez and Brian Granger for their technical advice.

References

Afgan E, Baker D, Van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, Grüning B. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic acids research. 2016:p.gkw343.

Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nature genetics. 2006; 38(5):500–501. [PubMed: 16642009]

Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B. KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter. 2009; 11(1):26–31.

Knuth DE. Literate programming. The Computer Journal. 1984; 27(2):97–111.

Leisch, F. Compstat. Physica-Verlag HD; 2002. Sweave: Dynamic generation of statistical reports using literate data analysis; p. 575-580.

Ragan-Kelley, et al. The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication. AGU Fall Meeting Abstracts. 2014 Dec. 1:07.

Highlights

- We integrated the GenePattern genomics platform with the Jupyter Notebook environment
- Notebooks interleave text, graphics, and analyses into complete “research narratives”
- Users can embed genomic analyses into notebooks without the need to write code
- GenePattern Notebook is freely available at http://www.genepattern-notebook.org

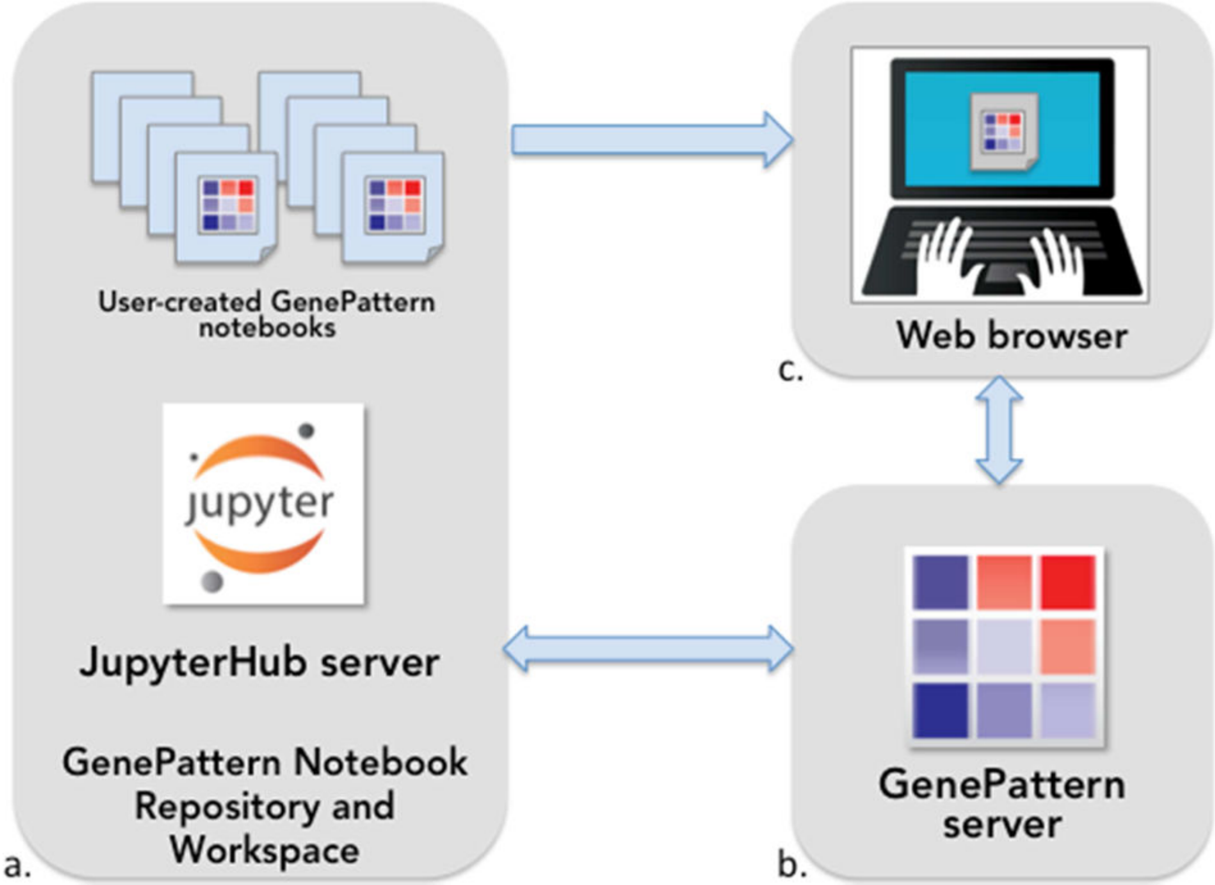

User-created GenePattern notebooks
jupyter
JupyterHub server
GenePattern Notebook Repository and Workspace
a.
Web browser
c.
GenePattern server
b.

Fig 1.

The GenePattern Notebook environment consists of a) an online environment, powered by JupyterHub, where users can create, share, and publish GenePattern notebooks; b) a GenePattern server that provides hundreds of pre-packaged genomic and machine learning analyses, all accessible through c) a Web browser.