Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace

Kun Qu[1,12], Sara Garamszegi[2,12], Felix Wu[2,12], Helga Thorvaldsdottir[2], Ted Liefeld[2,3], Marco Ocana[2,3], Diego Borges-Rivera[4], Nathalie Pochet[2,5], James T Robinson[2,3], Barry Demchak[3], Tim Hull[3], Gil Ben-Artzi[6,7], Daniel Blankenberg[8], Galt P Barber[9], Brian T Lee[9], Robert M Kuhn[9], Anton Nekrutenko[8], Eran Segal[6], Trey Ideker[3], Michael Reich[2,3], Aviv Regev[2,4,10], Howard Y Chang[1,11] & Jill P Mesirov[2,3]

Complex biomedical analyses require the use of multiple software tools in concert and remain challenging for much of the biomedical research community. We introduce GenomeSpace (http://www.genomespace.org), a cloud-based, cooperative community resource that currently supports the streamlined interaction of 20 bioinformatics tools and data resources. To facilitate integrative analysis by non-programmers, it offers a growing set of 'recipes', short workflows to guide investigators through high-utility analysis tasks.

© 2016 Nature America, Inc. All rights reserved.

The integrative analysis of diverse data types with multiple software tools remains an enormous challenge for many biologists. There is an ever-growing gap between the need to use various analysis and visualization tools and the complications of getting tools from different sources to work together. Moreover, it is difficult—even for experts, but especially for less computationally oriented biologists—to keep up with all of the available tools and to identify the right recipes to follow, particularly in the absence of an accepted 'laboratory manual' for analytic protocols.

Here, we present GenomeSpace, an open-source interoperability platform and community resource to enable non-programming scientists to work easily across data types and analysis methods (http://www.genomespace.org). GenomeSpace provides a 'tool launch pad' into which tools can be seamlessly added, and a 'data highway' that handles transfers between tools through format converters, relieving scientists of the burden of identifying and scripting the conversions. The GenomeSpace Recipe Resource is a growing set of high-utility use cases that demonstrate how to leverage multiple tools and serve as quick guides to analysis tasks. The website serves as a knowledge base, newsstand and point of contact and help for the community of users and tool developers.

Initially seeded by a consortium of biology research labs and development teams of six popular bioinformatics tools (Cytoscape[1,2], Galaxy[3], GenePattern[4], Genomica[5], the Integrative Genomics Viewer (IGV)[6] and the UCSC Table Browser[7]), GenomeSpace now connects 20 tools and data resources. Our consortium labs provided biological projects and analytical needs to drive GenomeSpace design and development. For example, we recapitulated the steps and results of published analyses[8,9] within GenomeSpace (**Supplementary Figs. 1** and **2**), dissecting and visualizing the gene regulatory networks in human cancer stem cells (**Supplementary Note 1**, **Supplementary Figs. 2–5**). The study required diverse data types, analytical steps and methods and multiple data transfers between tools. While originally requiring substantial scripting, this work can now be performed by non-programming biologists using only the GenomeSpace platform and tools within it.

From a user's perspective (**Fig. 1** and **Supplementary Fig. 6**), GenomeSpace has several features that together facilitate integrative analysis with a low barrier to user entry: (i) the collection of resident tools spanning a broad range of applications (**Table 1**); (ii) easy dataset management in a variety of cloud storage types, alongside data-sharing capabilities (all account holders receive an allocation of cloud storage, and GenomeSpace also supports connections to other cloud accounts (Dropbox, Google Drive, Amazon S3)); (iii) the ability to launch tools and to move data and analyses between tools, all facilitated by 'behind-the-scenes' file format converters (each tool retains its native environment and presents the same user interface and functionality as when launched outside of GenomeSpace); and (iv) a lightweight, simple, unifying web interface. In summary, from the web interface a researcher can launch a tool and simultaneously feed it input data files, move analysis results into other tools as needed through simple launching operations, and collect additional processed data within cloud accounts or local storage.

We developed the GenomeSpace Recipe Resource (http://www.genomespace.org/recipes) to aid researchers in identifying the steps required to perform a genomic analysis—a challenging task

[1]Program in Epithelial Biology, Stanford University School of Medicine, Stanford, California, USA. [2]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Department of Medicine, University of California, San Diego, La Jolla, California, USA. [4]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Program in Translational NeuroPsychiatric Genomics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [6]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [7]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [8]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania, USA. [9]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, California, USA. [10]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [11]Howard Hughes Medical Institute, Stanford University, Stanford, California, USA. [12]These authors contributed equally to this work. Correspondence should be addressed to J.P.M. (jmesirov@ucsd.edu).

RECEIVED 15 APRIL 2015; ACCEPTED 4 DECEMBER 2015; PUBLISHED ONLINE 18 JANUARY 2016; DOI:10.1038/NMETH.3732

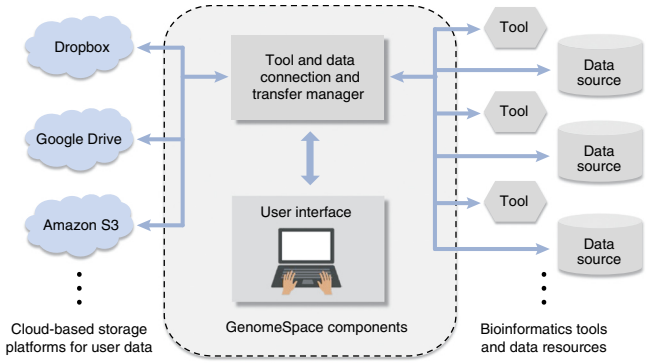

Figure 1 | The GenomeSpace environment for interoperation of bioinformatics tools.

even for short analyses. Although preconstructed pipelines can embody the entire workflow of a study, they may be insufficiently flexible for exploratory research. We took an alternative approach by providing a collection or 'cookbook' of recipes: i.e., comprehensive descriptions of cross-tool analysis workflows. Recipes are generally short, involving two or three tools, but commoditize important research tasks that investigators can employ as part of more complex analyses. The notion of our Recipe Resource is modeled after the classical lab guide *Molecular Cloning: A Laboratory Manual*[10], which used a similar approach to democratize molecular biology three decades ago.

Each GenomeSpace recipe contains a motivating biological problem, a relevant example dataset, detailed recipe steps and one possible interpretation of the results. Recipes include screenshot guides and videos that help walk users through the workflow. Current recipes cover diverse genomic analyses as well as basic utilities for using GenomeSpace itself (**Supplementary Table 1**). We are adding social media vehicles to make recipe collection a crowdsourced, collaborative effort, and we encourage suggestions for new multitool recipes and ideas to improve existing recipes.

An example from the Recipe Resource illustrates how to "Find subnetworks of differentially expressed genes and identify associated biological functions." Briefly, given a gene expression dataset, this recipe identifies network interactions between differentially expressed genes and annotates the biological functions within subnetworks via the Gene Ontology (GO) (**Supplementary Fig. 7**). The example dataset provided with this recipe is gene expression data from a study in which granulocyte-macrophage progenitor cells were transformed into leukemia stem cells by the introduction of an oncogene, *MLL-AF9* (ref. 11). Applying the recipe identifies processes that are correlated with transformation from a normal to a leukemic phenotype (**Supplementary Fig. 8**), such as *SMAD1*-dependent signaling, a process associated with the regulation of hematopoietic differentiation by *TGF-β* and *BMP*[12]. We also describe a second recipe example, "Identify biological functions for genes in copy number variation (CNV) regions" (**Supplementary Note 2** and **Supplementary Figs. 9** and **10**).

An important design goal was to facilitate rapid addition of diverse tools contributed by the developer community. This provides mutual benefit by extending the capabilities of GenomeSpace while also giving independent developers' tools simultaneous access to all GenomeSpace-connected tools and data sources, circumventing the need to connect to each one individually. Recent cross-tool interoperability efforts have used one of several approaches: aggregators host a large number of command line tools (Galaxy, GenePattern); plug-in architectures

Table 1 | GenomeSpace provides access to a diverse set of bioinformatics tools and resources

Tool name	Organization	Project website
Analysis and visualization tools		
Cistrome	Dana-Farber Cancer Institute	http://www.cistrome.org
Cytoscape 3[a]	Cytoscape Consortium	http://www.cytoscape.org
Cytoscape 2[a]	Cytoscape Consortium	http://www.cytoscape.org
Galaxy	Pennsylvania State University and Johns Hopkins University	http://www.galaxyproject.org
GenePattern	Broad Institute and University of California, San Diego (UCSD)	http://www.genepattern.org
Genomica	Weizmann Institute of Science	http://genomica.weizmann.ac.il
geWorkbench	Columbia University	http://www.geworkbench.org
Gitools	University Pompeu Fabra, Barcelona	http://www.gitools.org
Integrative Genomics Viewer (IGV)	Broad Institute and UCSD	http://www.igv.org
ISAcreator	University of Oxford	http://www.isa-tools.org
Molecular Signatures Database (MSigDB) Online Tools	Broad Institute and UCSD	http://www.msigdb.org
Data resources		
ArrayExpress	European Bioinformatics Institute	http://www.ebi.ac.uk/arrayexpress
InSilicoDB	InSilico Genomics	http://insilicodb.com
Synapse	Sage Bionetworks	http://synapse.org
UCSC Table Browser	University of California Santa Cruz	http://genome.ucsc.edu
Integrated portals (data and analysis)		
Project Achilles	Dana-Farber Cancer Institute and Broad Institute	http://broadinstitute.org/achilles
Cancer Cell Line Encyclopedia (CCLE)	Broad Institute and Novartis Institutes for Biomedical Research	http://broadinstitute.org/ccle
cBioPortal for Cancer Genomics	Memorial Sloan Kettering Cancer Center	http://www.cbioportal.org
Multiple Myeloma Genomics Portal (MMGP)	Multiple Myeloma Research Consortium, Broad Institute and Translational Genomics Research Institute (TGen)	http://broadinstitute.org/mmgp
Reactome	Ontario Institute for Cancer Research, European Bioinformatics Institute and New York University Medical Center	http://www.reactome.org

[a]Cytoscape 3 and Cytoscape 2 have different underlying architectures and different user interfaces. Both versions are made available through GenomeSpace to accommodate users who may prefer one to the other.

© 2016 Nature America, Inc. All rights reserved.

provide a way to extend the functionality of a basic package (Cytoscape, geWorkbench[13], MeV[14]); and messaging systems send data and instructions between tools (MeDICi[15], Gaggle[16]). Our open-source, lightweight, hybrid approach combines aspects of both messaging and aggregating systems. The resulting platform (**Supplementary Fig. 11** and Online Methods) provides single sign-on for GenomeSpace tools and data resources; security mechanisms and user-controlled levels of sharing; and a common interface to multiple cloud storage providers. Moreover, this approach supports interoperation among diverse desktop and web-based tools, while minimizing the amount of effort required to connect to the platform (Online Methods).

To further facilitate cross-tool interoperability, GenomeSpace offers a range of file converters for directly converting between pairs of file formats (**Supplementary Note 3**). Direct conversion obviates the development burden of defining and supporting central data models for tools, especially legacy ones, connecting to GenomeSpace. Moreover, because converters are independent and do not rely on a GenomeSpace-specific data model, we can expand the set of supported formats by leveraging converters that are developed outside of GenomeSpace.

GenomeSpace thus provides several benefits that help ease analysis and expand the universe of options accessible to biologists. First, it allows seamless transition between tools. Automatic file format converters speed tasks such as launching and moving data between tools and obviate the need for custom conversion scripts, an insurmountable barrier for many biologists. Second, the large set of connected tools allows data to be examined in greater depth and diversity than with any single tool, enriching the interpretation of integrative analyses. Third, we encourage the inclusion of multiple tools with similar capabilities so that investigators can choose the tool with which they are most familiar and use alternatives to test their findings for robustness and reproducibility. Fourth, recipes provide short workflows that can be assembled into more complex analysis scenarios and can also introduce investigators to new analysis methods and tools.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank other members of the GenomeSpace and GenePattern Teams for their contributions and input: P. Carr, B. Hill-Meyers, S.H. Lee and T. Tabor (Broad Institute of MIT and Harvard); J. Zhang (Stanford University); and H. Carter and M. Smoot (University of California, San Diego). Special thanks to D. Haussler and J. Kent (University of California, Santa Cruz) for their involvement in the nascent stages of the GenomeSpace project. We thank J. Bistline for help with the citations and figures, and L. Gaffney for help with the figures. This work has been supported by US National Institutes of Health–National Human Genome Research Institute P01 HG005062 and U41 HG007517, with additional initial support from Amazon Web Services (AWS).

AUTHOR CONTRIBUTIONS

M.R., A.R. and J.P.M. conceived of the GenomeSpace concept. T.L., M.O. and M.R. designed and implemented the GenomeSpace software. K.Q., S.G., F.W. and N.P. implemented the driving biological projects within GenomeSpace with supervision and input from A.R., H.Y.C. and J.P.M. The recipes were implemented by S.G., F.W. and D.B.-R. The GenomeSpace seed tools were added to the system by J.T.R., B.D., T.H., G.B.-A., D.B., G.P.B., B.T.L., R.M.K., A.N., E.S. and T.I., who also consulted on the GenomeSpace architecture. H.T., M.R., A.R., H.Y.C. and J.P.M. supervised the GenomeSpace project. K.Q., S.G., F.W., H.T., M.R., H.C., A.R. and J.P.M. wrote the manuscript. All authors reviewed and approved the final manuscript as submitted.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Demchak, B. *et al. F1000Res.* **3**, 151 (2014).
2. Shannon, P. *et al. Genome Res.* **13**, 2498–2504 (2003).
3. Giardine, B. *et al. Genome Res.* **15**, 1451–1455 (2005).
4. Reich, M. *et al. Nat. Genet.* **38**, 500–501 (2006).
5. Segal, E., Friedman, N., Koller, D. & Regev, A. *Nat. Genet.* **36**, 1090–1098 (2004).
6. Robinson, J.T. *et al. Nat. Biotechnol.* **29**, 24–26 (2011).
7. Karolchik, D. *et al. Nucleic Acids Res.* **32**, D493–D496 (2004).
8. Ben-Porath, I. *et al. Nat. Genet.* **40**, 499–507 (2008).
9. Wong, D.J. *et al. Cell Stem Cell* **2**, 333–344 (2008).
10. Sambrook, J., Fritsch, E.F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* vol. 3 (Cold Spring Harbor Laboratory Press, 1989).
11. Krivtsov, A.V. *et al. Nature* **442**, 818–822 (2006).
12. Larsson, J. & Karlsson, S. *Oncogene* **24**, 5676–5692 (2005).
13. Floratos, A., Smith, K., Ji, Z., Watkinson, J. & Califano, A. *Bioinformatics* **26**, 1779–1780 (2010).
14. Saeed, A.I. *et al. Biotechniques* **34**, 374–378 (2003).
15. Gorton, I., Wynne, A., Almquist, J. & Chatterton, J. in *Software Architecture, 2008. WICSA 2008. Seventh Working IEEE/IFIP Conference* (eds. Kruchten, P., Garlan, D. & Woods, E.) 95–104 (IEEE Computer Society, 2008).
16. Shannon, P.T., Reiss, D.J., Bonneau, R. & Baliga, N.S. *BMC Bioinformatics* **7**, 176 (2006).

© 2016 Nature America, Inc. All rights reserved.

ONLINE METHODS

GenomeSpace architecture. GenomeSpace presents a 'connection layer' that includes a collection of web services with well-defined entry points to the GenomeSpace server that provides the core system functionality (**Supplementary Fig. 11**). The web user interface also interacts with the server through these entry points. The GenomeSpace server currently runs as an Amazon Machine Instance (AMI) in the Amazon Elastic Compute Cloud (EC2). It consists of three components. (1) An Identity Service manages sign-on credentials, including single sign-on to the GenomeSpace tools, and data access. GenomeSpace leverages the Amazon AWS security mechanisms, which are compliant with the requirements from many standards organizations and government agencies. All data are private by default, but users may share directories or files with other users or groups of users. (2) An Analysis Tools Manager maintains information about tool capabilities and dependencies and coordinates tool launches, including the ability to launch other GenomeSpace tools from within a tool; (3) A Data Manager handles data storage, transfer to/from the cloud (including Amazon S3, Google Drive, and Dropbox), data sharing, and the file format conversions that provide a smooth script-free connection between tools.

Connecting tools to GenomeSpace. The GenomeSpace connection layer includes a collection of web services with well-defined entry points to the GenomeSpace server that provides the core system functionality. It is available as Java and JavaScript client development kits for tools developed in those languages, or as web services with a RESTful application programming interface (API) for any language. Developers can also take advantage of a number of user interface widgets that are available for common user tasks, including file chooser dialogs and authentication panels. Adding a tool to GenomeSpace, using these resources, typically takes on the order of two programmer days or less, depending on the type of tool. The most recent tool to join the community was cBioPortal (http://www.cbioportal.org) from Memorial Sloan Kettering Cancer Center, and the development team reported that it took an hour to connect this web-based portal as a data source to GenomeSpace. We note that command-line tools that do not have their own user interface can join the GenomeSpace community via either of its current aggregator members—GenePattern and Galaxy.

GenomeSpace source code. The GenomeSpace source code is available on http://bitbucket.org/GenomeSpace/combined, under the GNU LGPL version 2.1 license. Technical resources for software developers are available on the GenomeSpace website at http://genomespace.org/for-developers/.

© 2016 Nature America, Inc. All rights reserved.

doi:10.1038/nmeth.3732