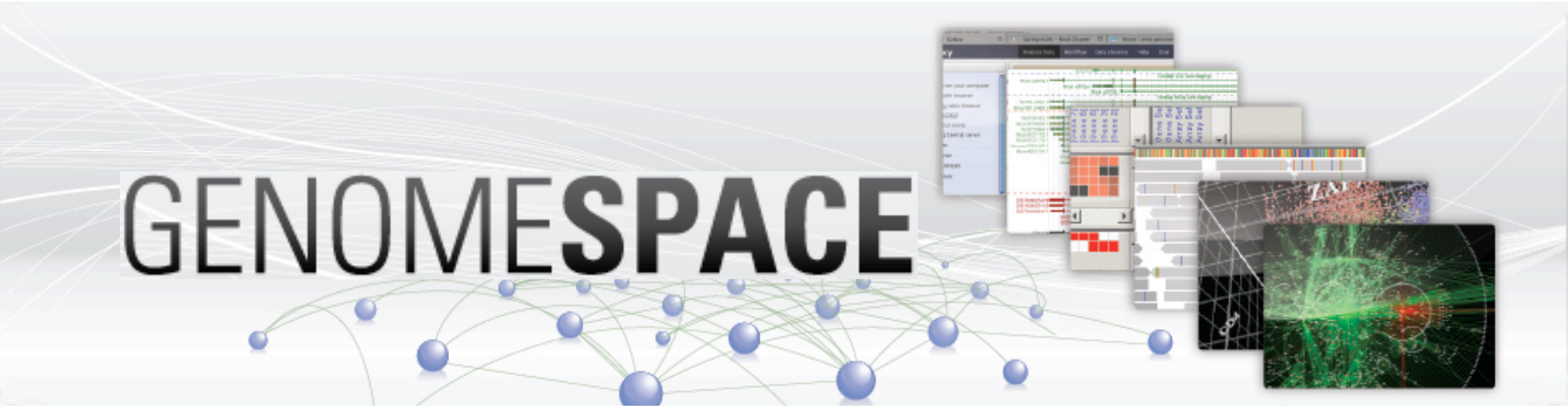


(tools x connections) + recipes = frictionless science

Michael Reich

February 5, 2016

Interoperable, integrative computational genomics

Need

Insights through integrative studies of all data types

Challenges

Flood of high throughput data

- genomic sequence
- global mRNA expression profiles
- copy number and LOH
- epigenetic data
- protein level and modification status
- metabolite profiles

Proliferation of tools

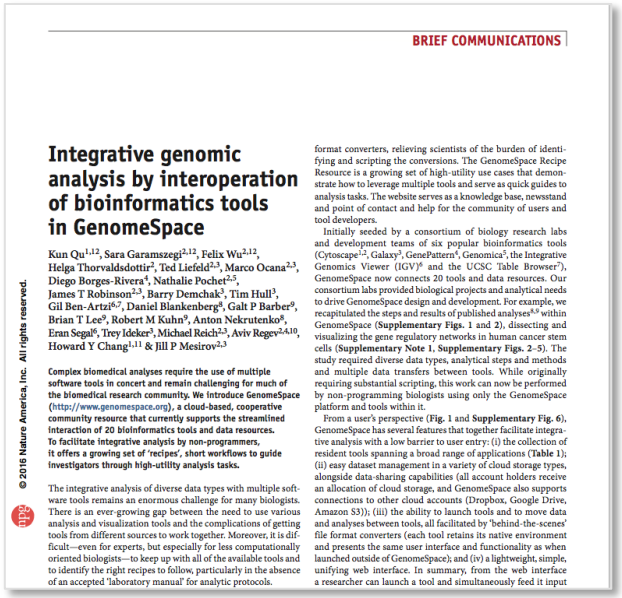
- Databases, visualization, and analysis
- Difficulty of getting tools to work together
- Access, analyze, visualize each data type separately

GENOMESPACE

An online community to share computational tools

www.genomespace.org

- 12,000 registered users
- Published in Nature Methods: Ku et al, 2016



BRIEF COMMUNICATIONS

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[2], 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.

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.

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

© 2016 Nature America, Inc. All rights reserved.

GenomeSpace:
Interoperable, Integrative Genomic Analysis

Connect Collaborate Combine

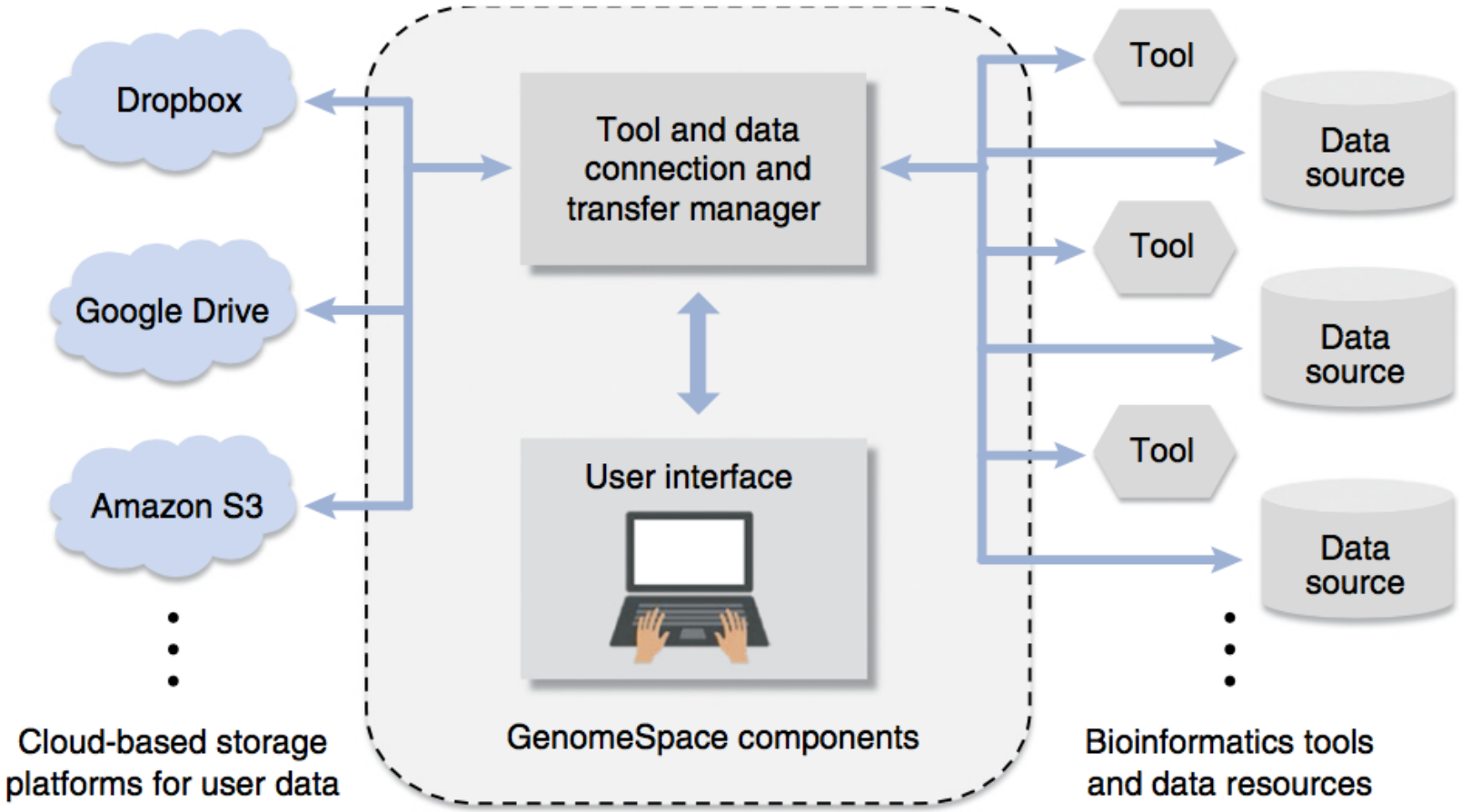
Lightweight "connection layer" between tools and data sources, with automatic format conversion.

Cloud-based storage makes tools and data accessible from any location.

Dropbox Google Drive amazon web services

Tools retain their native look and feel.

GenomeSpace Architecture



GenomeSpace Tools 2/16

Trinity

ArrayExpress

Cancer Cell Line Encyclopedia

Cistrome

Cytoscape

Cytoscape 3

Galaxy

GenePattern

Genomica

geWorkbench

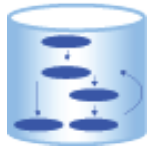
Gitools

IGV

InSilicoDB

ISACreator

Multiple Myeloma Genomics Portal



MSigDB

Project Achilles

Reactome

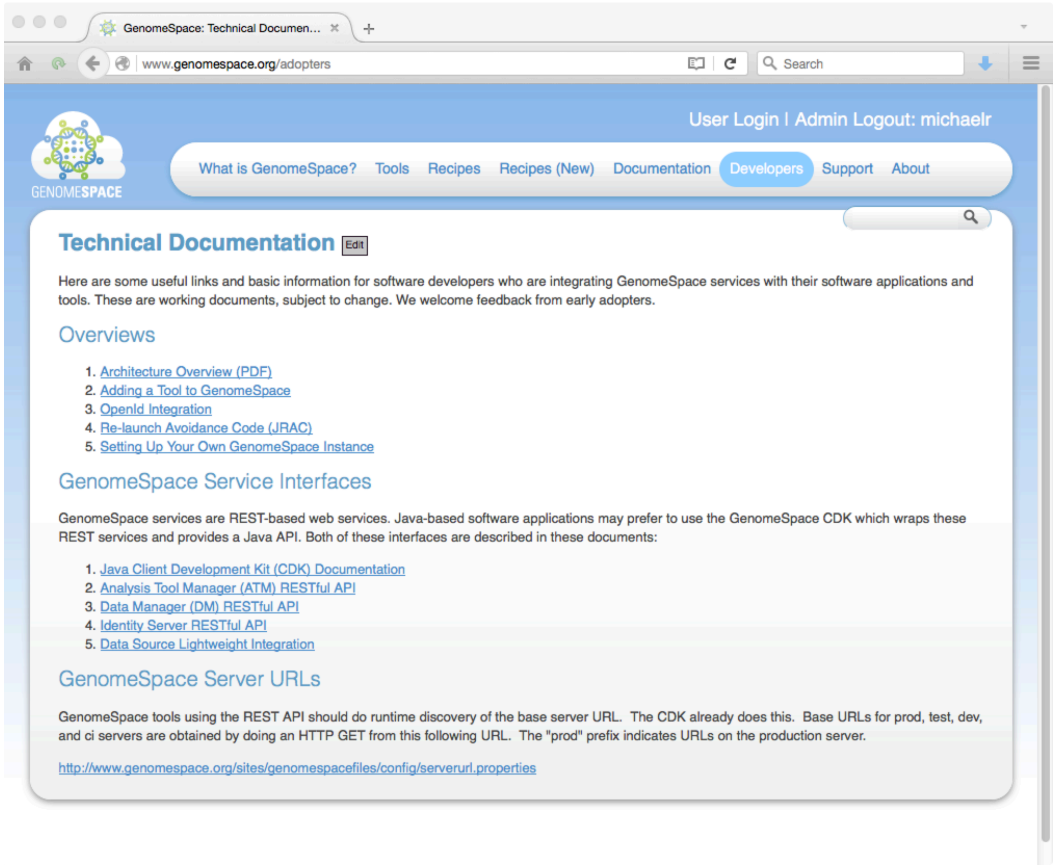
Synapse

UCSC Table Browser

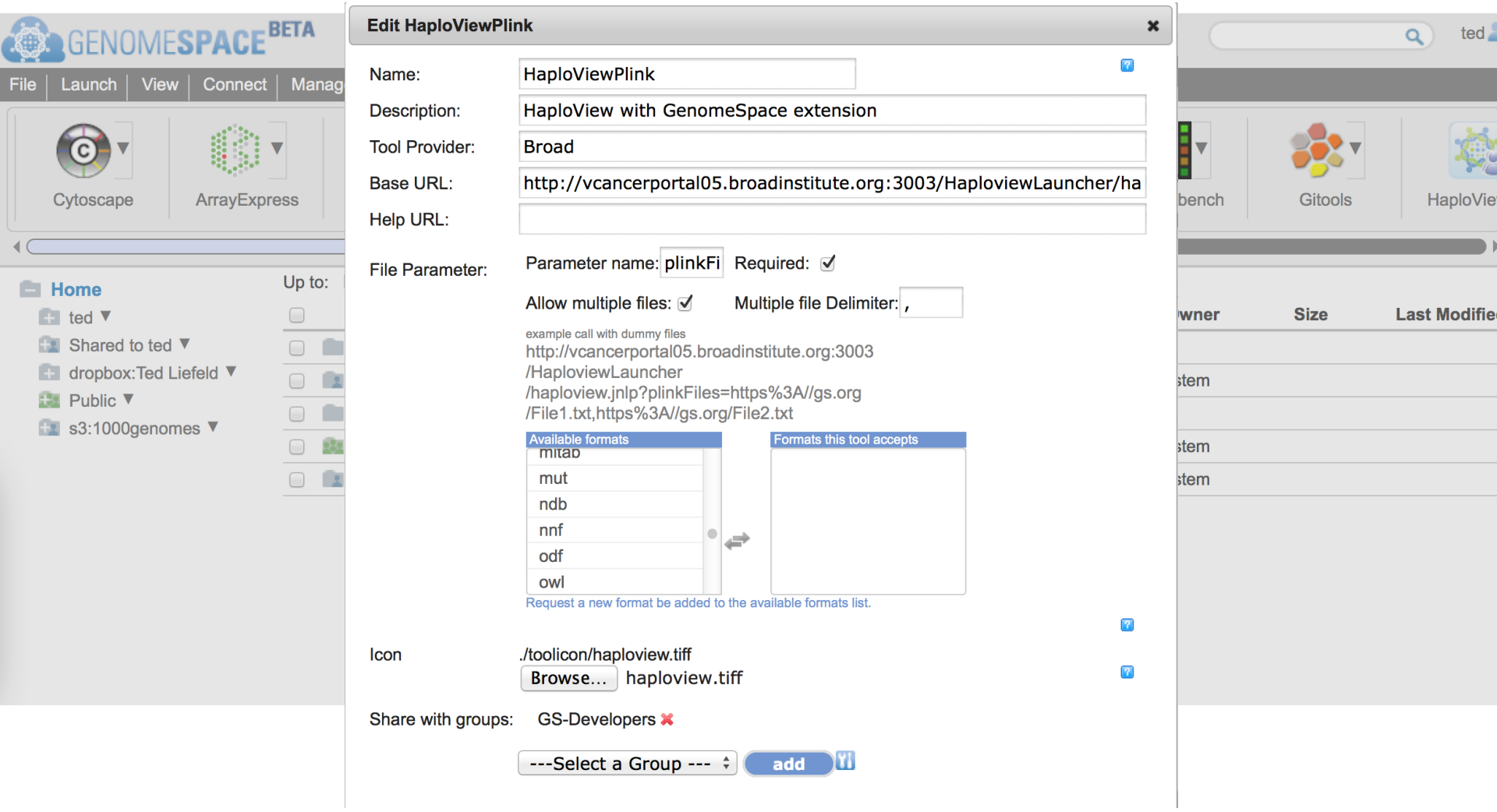
cBio Portal

GenomeSpace-enablement of new tools

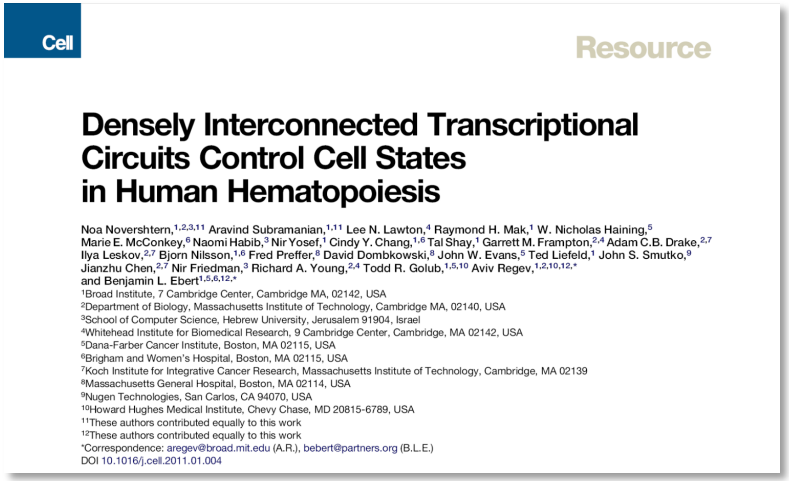
- Basic enablement in ≤ 1 programmer-day
- RESTful API
- Client Development Kit (CDK)
- Data Transfer Utilities
- Documentation



Add your own private tools



GenomeSpace in action: Identification of transcription factor regulators



Cell

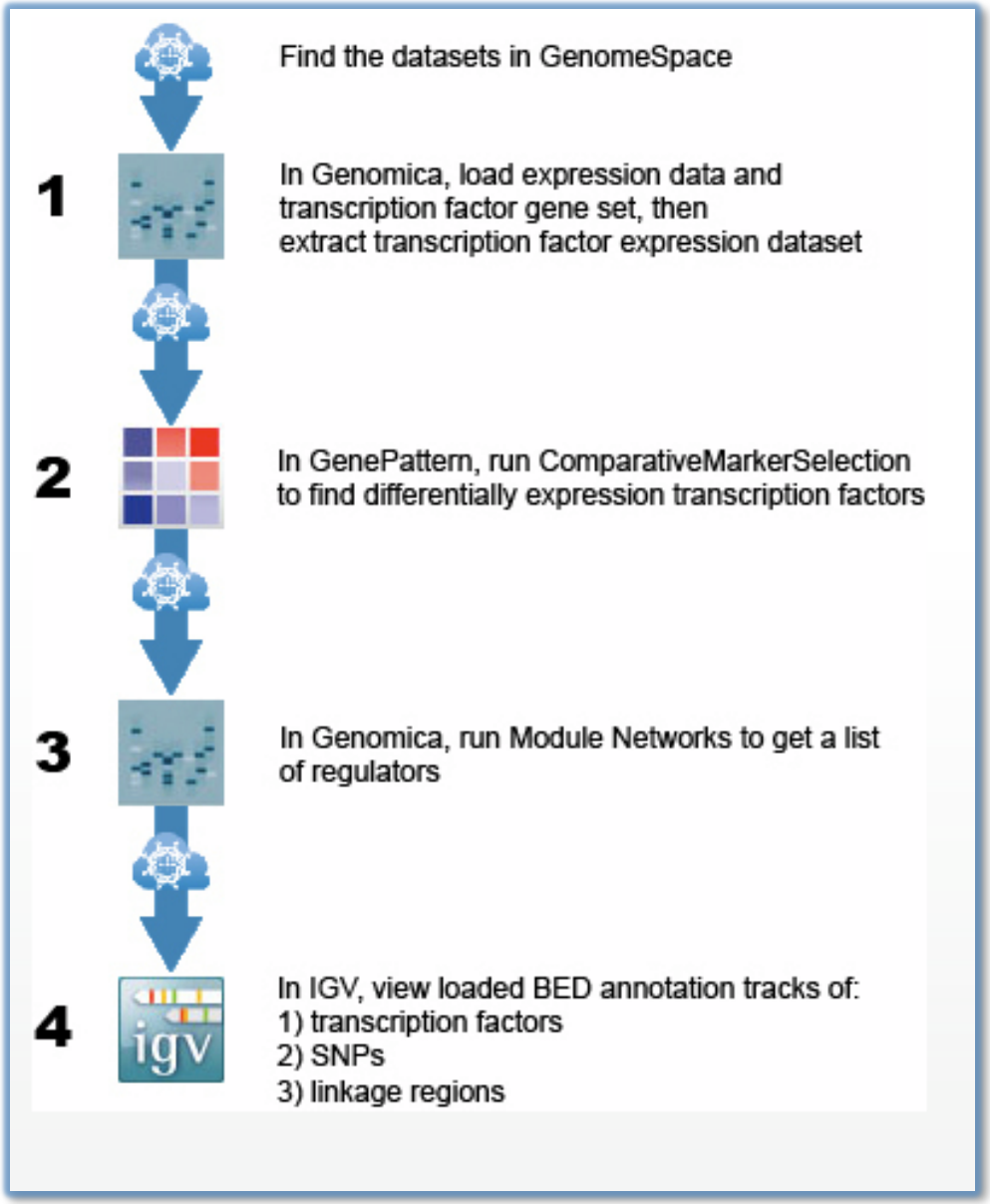
Resource

Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis

Noa Novershtern,[1,2,3,11] Aravind Subramanian,[1,11] Lee N. Lawton,[4] Raymond H. Mak,[1] W. Nicholas Haining,[5] Marie E. McConkey,[6] Naomi Habib,[3] Nir Yosef,[1] Cindy Y. Chang,[1,6] Tal Shay,[1] Garrett M. Frampton,[2,4] Adam C.B. Drake,[2,7] Ilya Leskov,[2,7] Bjorn Nilsson,[1,6] Fred Preffer,[8] David Dombkowski,[8] John W. Evans,[5] Ted Liefeld,[1] John S. Smutko,[9] Jianzhu Chen,[2,7] Nir Friedman,[3] Richard A. Young,[2,4] Todd R. Golub,[1,5,10] Aviv Regev,[1,2,10,12,*] and Benjamin L. Ebert[1,5,6,12,*]

[1]Broad Institute, 7 Cambridge Center, Cambridge MA, 02142, USA
[2]Department of Biology, Massachusetts Institute of Technology, Cambridge MA, 02140, USA
[3]School of Computer Science, Hebrew University, Jerusalem 91904, Israel
[4]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA
[5]Dana-Farber Cancer Institute, Boston, MA 02115, USA
[6]Brigham and Women's Hospital, Boston, MA 02115, USA
[7]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139
[8]Massachusetts General Hospital, Boston, MA 02114, USA
[9]Nugen Technologies, San Carlos, CA 94070, USA
[10]Howard Hughes Medical Institute, Chevy Chase, MD 20815-6789, USA
[11]These authors contributed equally to this work
[12]These authors contributed equally to this work
*Correspondence: aregev@broad.mit.edu (A.R.), bebert@partners.org (B.L.E.)
DOI 10.1016/j.cell.2011.01.004

Novershtern et al., Cell 2011

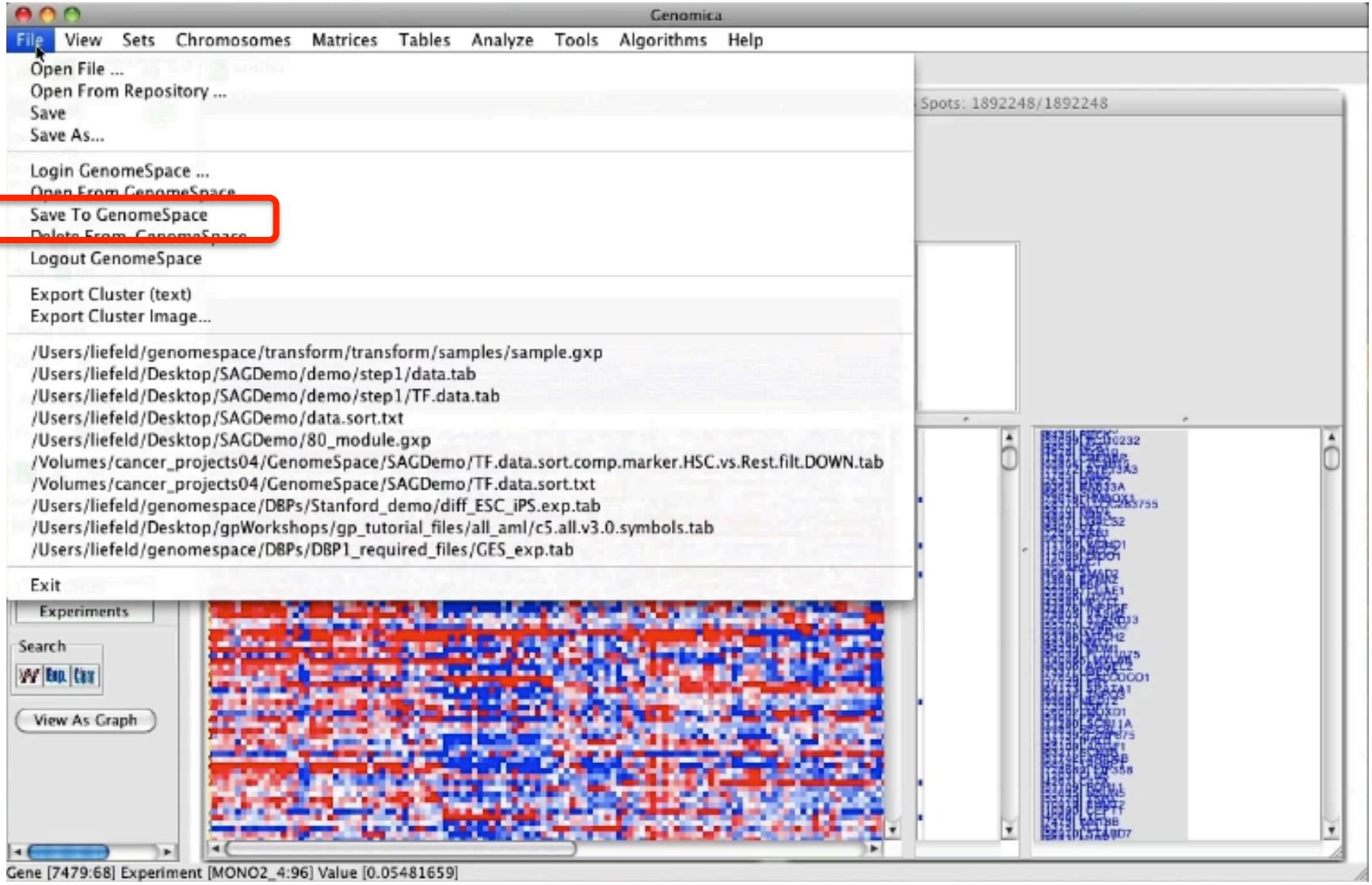


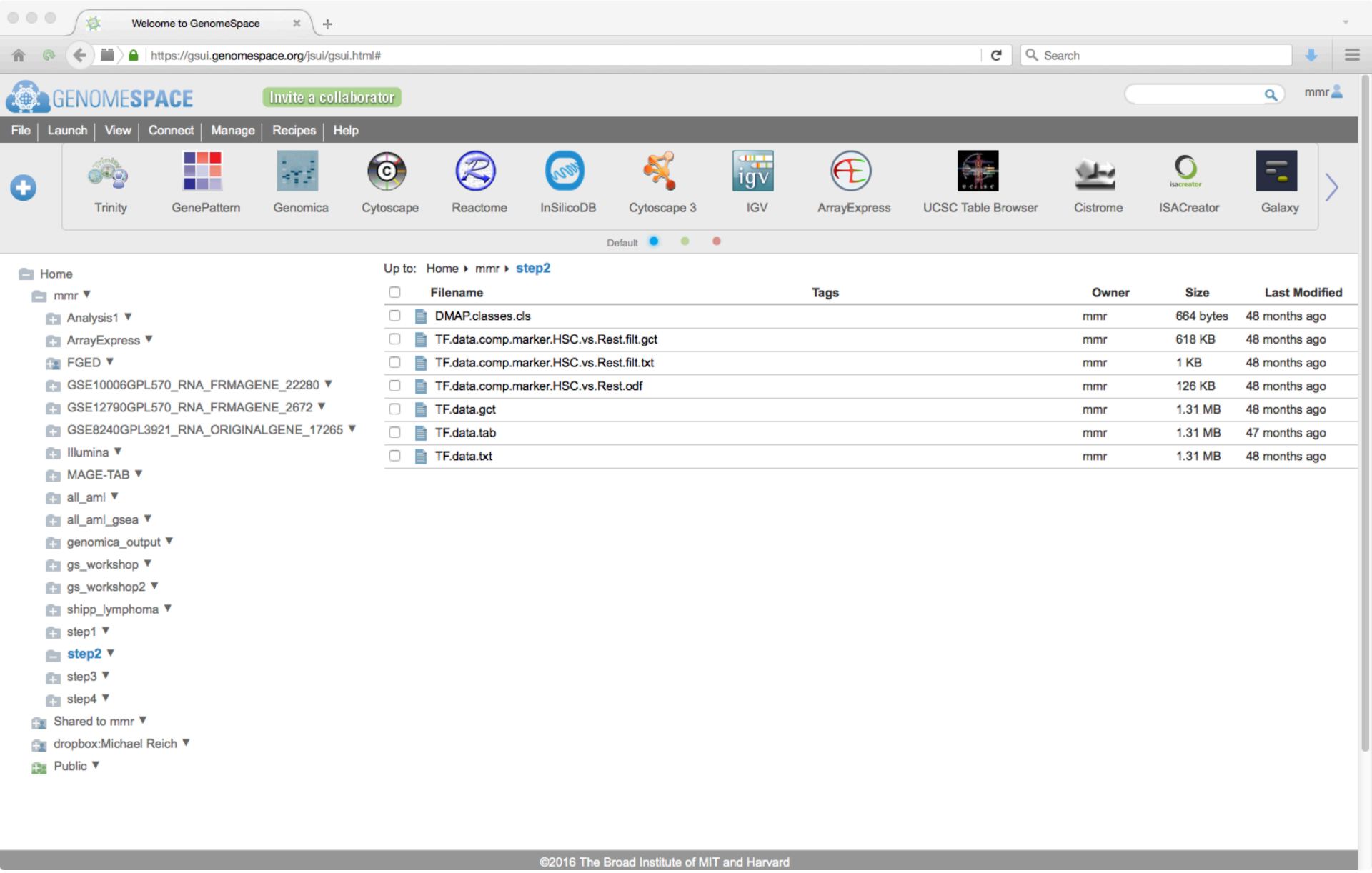
Doing this the standard way

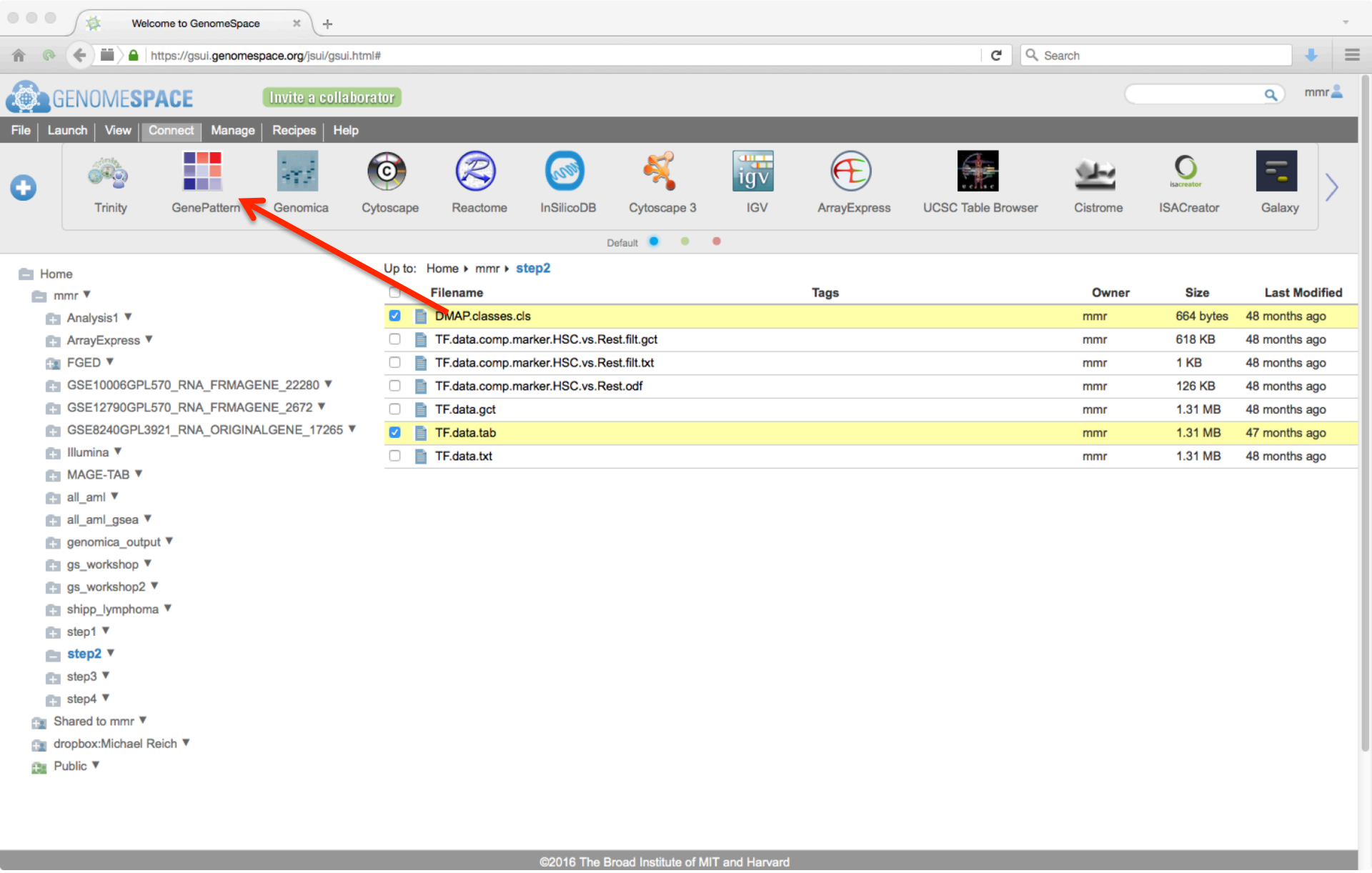
1. Load file into Genomica
2. Run Genomica analysis to create transcription factor dataset
3. Download result file
4. Convert result file to GenePattern format
5. Upload file to GenePattern
6. Run GenePattern differential expression analysis
7. Download result file
8. Convert result file back to Genomica format
9. Upload to Genomica
10. Perform module network analysis
11. Download result file
12. Convert result file to IGV format
13. Open IGV
14. Load Genomica results into IGV
15. Visualize results

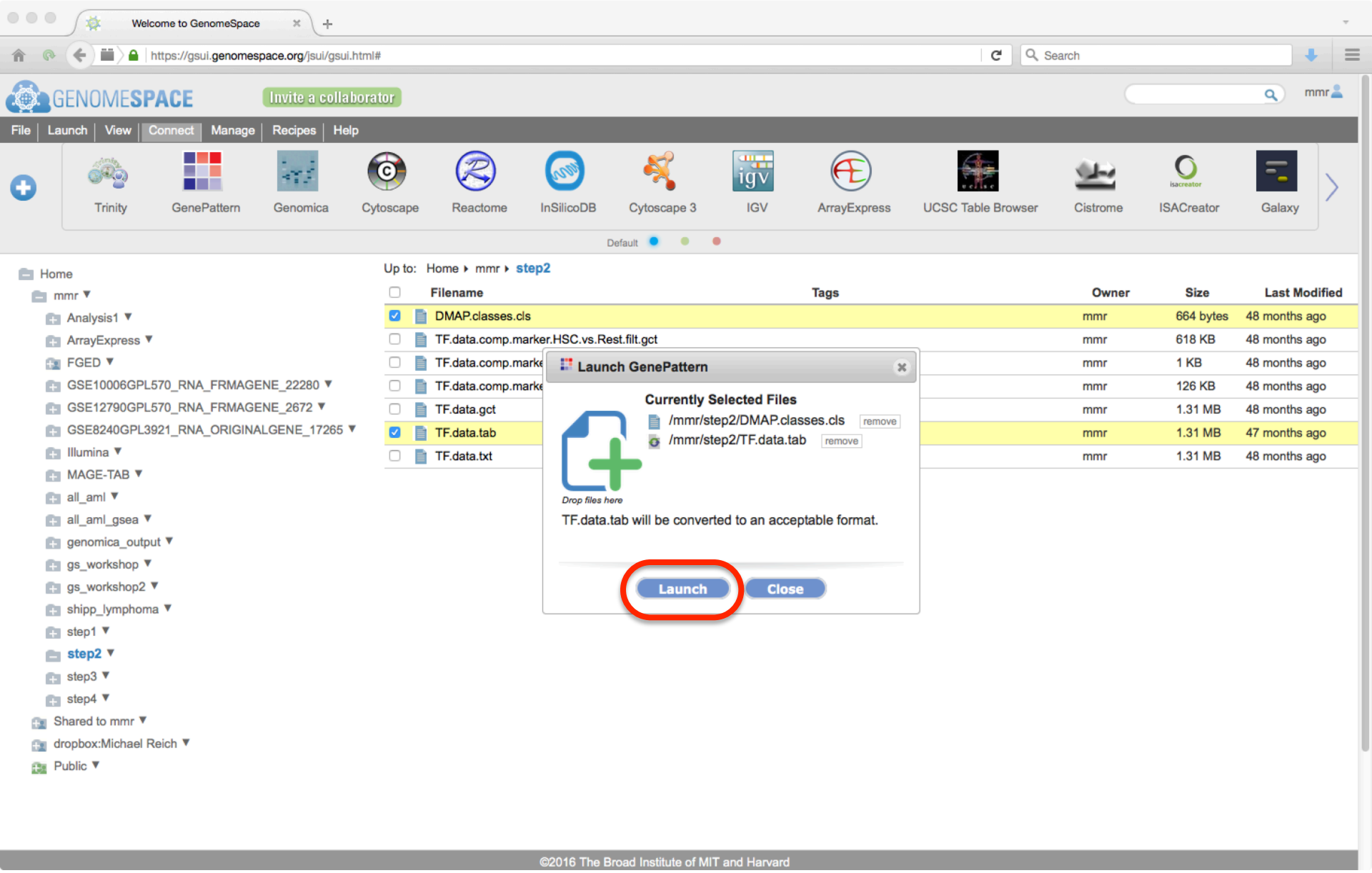
Script
Data transfer

Create transcription factor dataset in Genomica and save to GenomeSpace

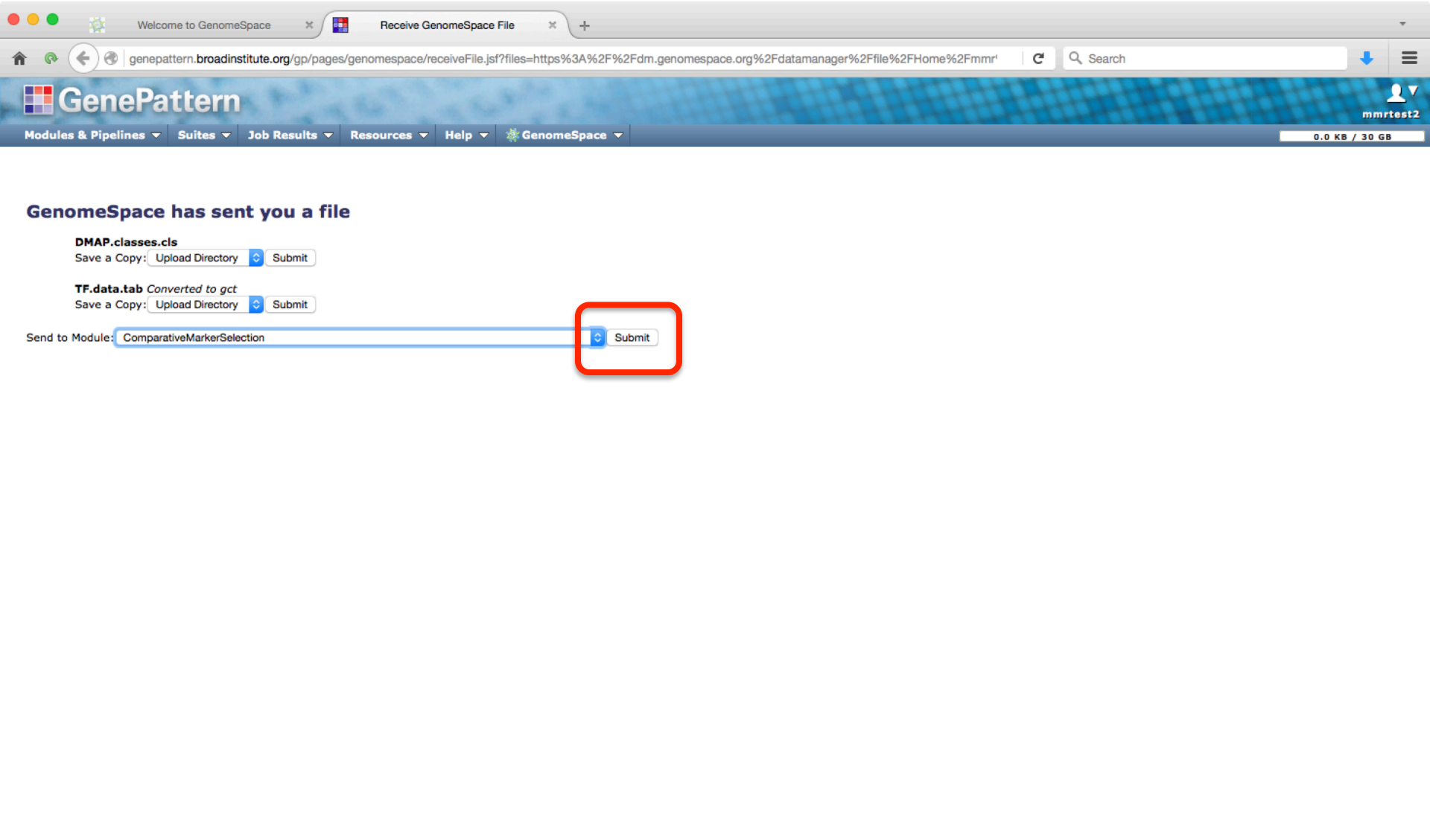



Welcome to GenomeSpace
https://gsui.genomespace.org/jsui/gsui.html#
Search
GENOMESPACE
Invite a collaborator
mmr
File
Launch
View
Connect
Manage
Recipes
Help
Trinity
GenePattern
Genomica
Cytoscape
Reactome
InSilicoDB
Cytoscape 3
IGV
ArrayExpress
UCSC Table Browser
Cistrome
ISACreator
Galaxy
Default
Home
mmr
Analysis1
ArrayExpress
FGED
GSE10006GPL570_RNA_FRMAGENE_22280
GSE12790GPL570_RNA_FRMAGENE_2672
GSE8240GPL3921_RNA_ORIGINALGENE_17265
Illumina
MAGE-TAB
all_aml
all_aml_gsea
genomica_output
gs_workshop
gs_workshop2
shipp_lymphoma
step1
step2
step3
step4
Shared to mmr
dropbox:Michael Reich
Public
Up to: Home ▸ mmr ▸ step2
Filename
Tags
Owner
Size
Last Modified
DMAP.classes.cls mmr 664 bytes 48 months ago
TF.data.comp.marker.HSC.vs.Rest.filt.gct mmr 618 KB 48 months ago
TF.data.comp.marker.HSC.vs.Rest.filt.txt mmr 1 KB 48 months ago
TF.data.comp.marker.HSC.vs.Rest.odf mmr 126 KB 48 months ago
TF.data.gct mmr 1.31 MB 48 months ago
TF.data.tab mmr 1.31 MB 47 months ago
TF.data.txt mmr 1.31 MB 48 months ago
©2016 The Broad Institute of MIT and Harvard

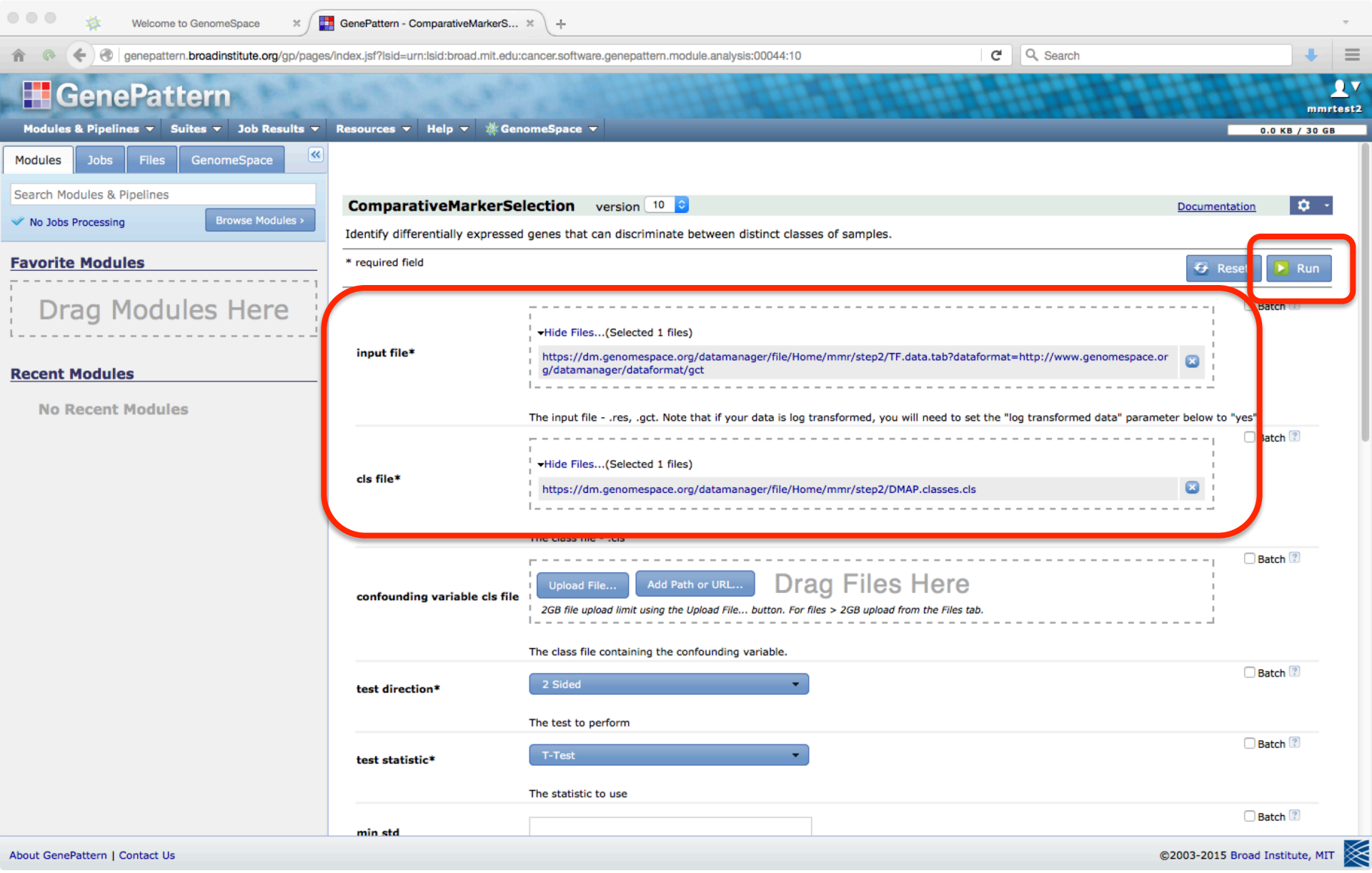

Welcome to GenomeSpace
https://gsui.genomespace.org/jsui/gsui.html#
Search
GENOMESPACE
Invite a collaborator
mmr
File
Launch
View
Connect
Manage
Recipes
Help
Trinity
GenePattern
Genomica
Cytoscape
Reactome
InSilicoDB
Cytoscape 3
IGV
ArrayExpress
UCSC Table Browser
Cistrome
ISACreator
Galaxy
Default
Home
mmr
Analysis1
ArrayExpress
FGED
GSE10006GPL570_RNA_FRMAGENE_22280
GSE12790GPL570_RNA_FRMAGENE_2672
GSE8240GPL3921_RNA_ORIGINALGENE_17265
Illumina
MAGE-TAB
all_aml
all_aml_gsea
genomica_output
gs_workshop
gs_workshop2
shipp_lymphoma
step1
step2
step3
step4
Shared to mmr
dropbox:Michael Reich
Public
Up to: Home ▸ mmr ▸ step2
Filename
Tags
Owner
Size
Last Modified
DMAP.classes.cls mmr 664 bytes 48 months ago
TF.data.comp.marker.HSC.vs.Rest.filt.gct mmr 618 KB 48 months ago
TF.data.comp.marker.HSC.vs.Rest.filt.txt mmr 1 KB 48 months ago
TF.data.comp.marker.HSC.vs.Rest.odf mmr 126 KB 48 months ago
TF.data.gct mmr 1.31 MB 48 months ago
TF.data.tab mmr 1.31 MB 47 months ago
TF.data.txt mmr 1.31 MB 48 months ago
©2016 The Broad Institute of MIT and Harvard


Welcome to GenomeSpace
https://gsui.genomespace.org/jsui/gsui.html#
Search
GENOMESPACE
Invite a collaborator
mmr
File
Launch
View
Connect
Manage
Recipes
Help
Trinity
GenePattern
Genomica
Cytoscape
Reactome
InSilicoDB
Cytoscape 3
IGV
ArrayExpress
UCSC Table Browser
Cistrome
ISACreator
Galaxy
Default
Home
mmr
Analysis1
ArrayExpress
FGED
GSE10006GPL570_RNA_FRMAGENE_22280
GSE12790GPL570_RNA_FRMAGENE_2672
GSE8240GPL3921_RNA_ORIGINALGENE_17265
Illumina
MAGE-TAB
all_aml
all_aml_gsea
genomica_output
gs_workshop
gs_workshop2
shipp_lymphoma
step1
step2
step3
step4
Shared to mmr
dropbox:Michael Reich
Public
Up to: Home ▸ mmr ▸ step2
Filename
Tags
Owner
Size
Last Modified
DMAP.classes.cls mmr 664 bytes 48 months ago
TF.data.comp.marker.HSC.vs.Rest.filt.gct mmr 618 KB 48 months ago
TF.data.comp.marke mmr 1 KB 48 months ago
TF.data.comp.marke mmr 126 KB 48 months ago
TF.data.gct mmr 1.31 MB 48 months ago
TF.data.tab mmr 1.31 MB 47 months ago
TF.data.txt mmr 1.31 MB 48 months ago
Launch GenePattern
Currently Selected Files
/mmr/step2/DMAP.classes.cls remove
/mmr/step2/TF.data.tab remove
Drop files here
TF.data.tab will be converted to an acceptable format.
Launch
Close
©2016 The Broad Institute of MIT and Harvard

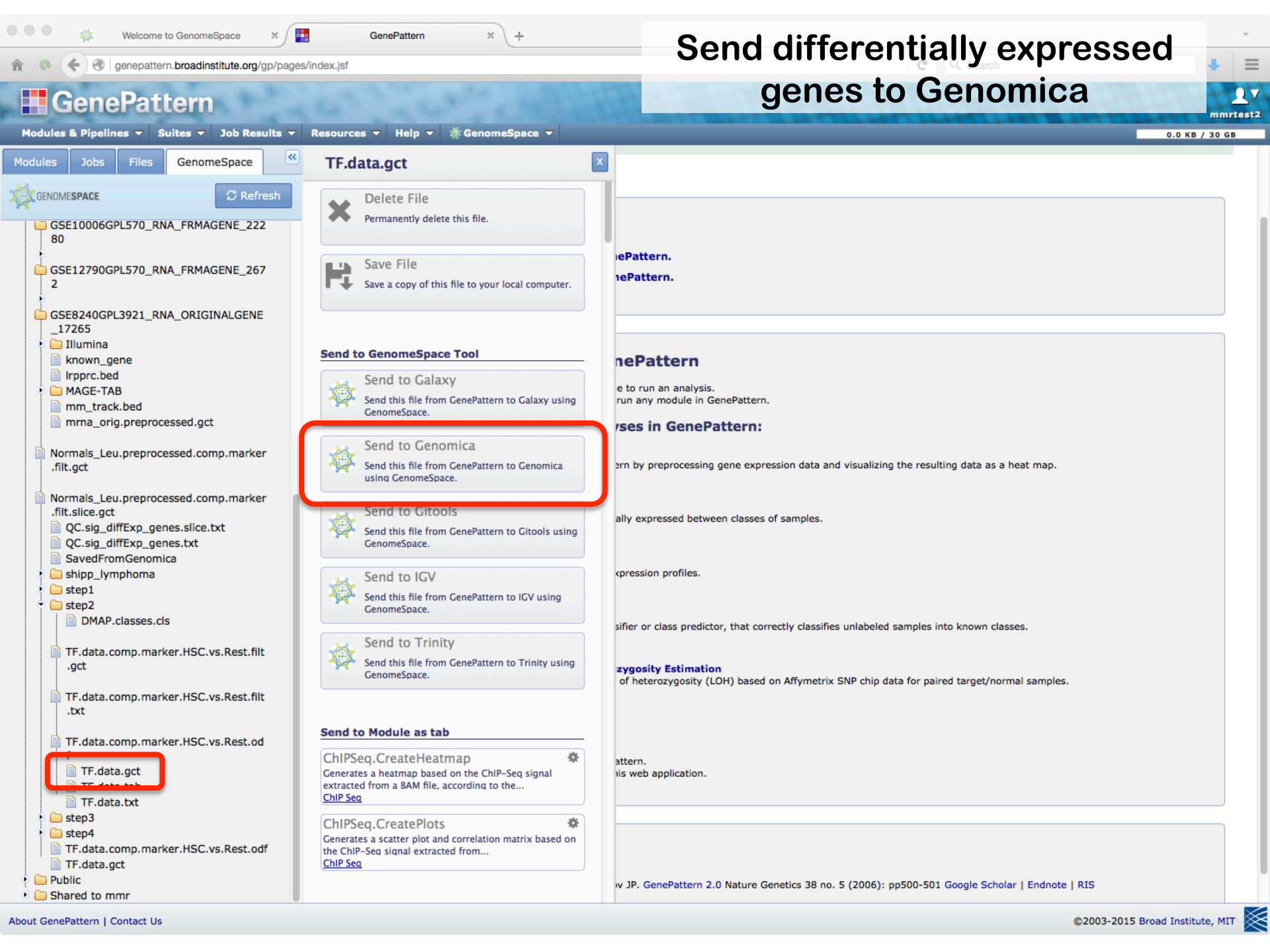
Perform differential expression analysis in GenePattern

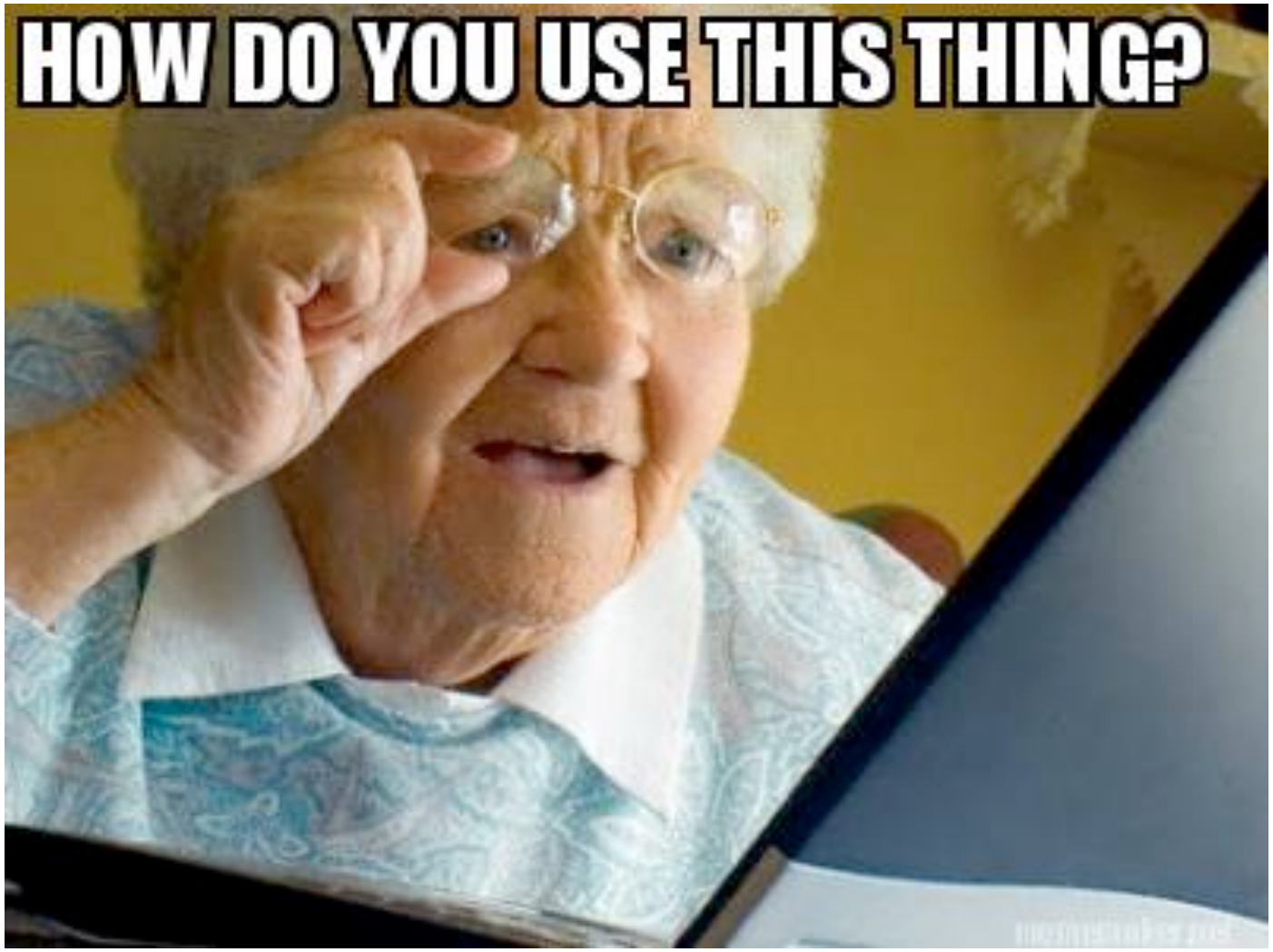


Perform differential expression analysis in GenePattern

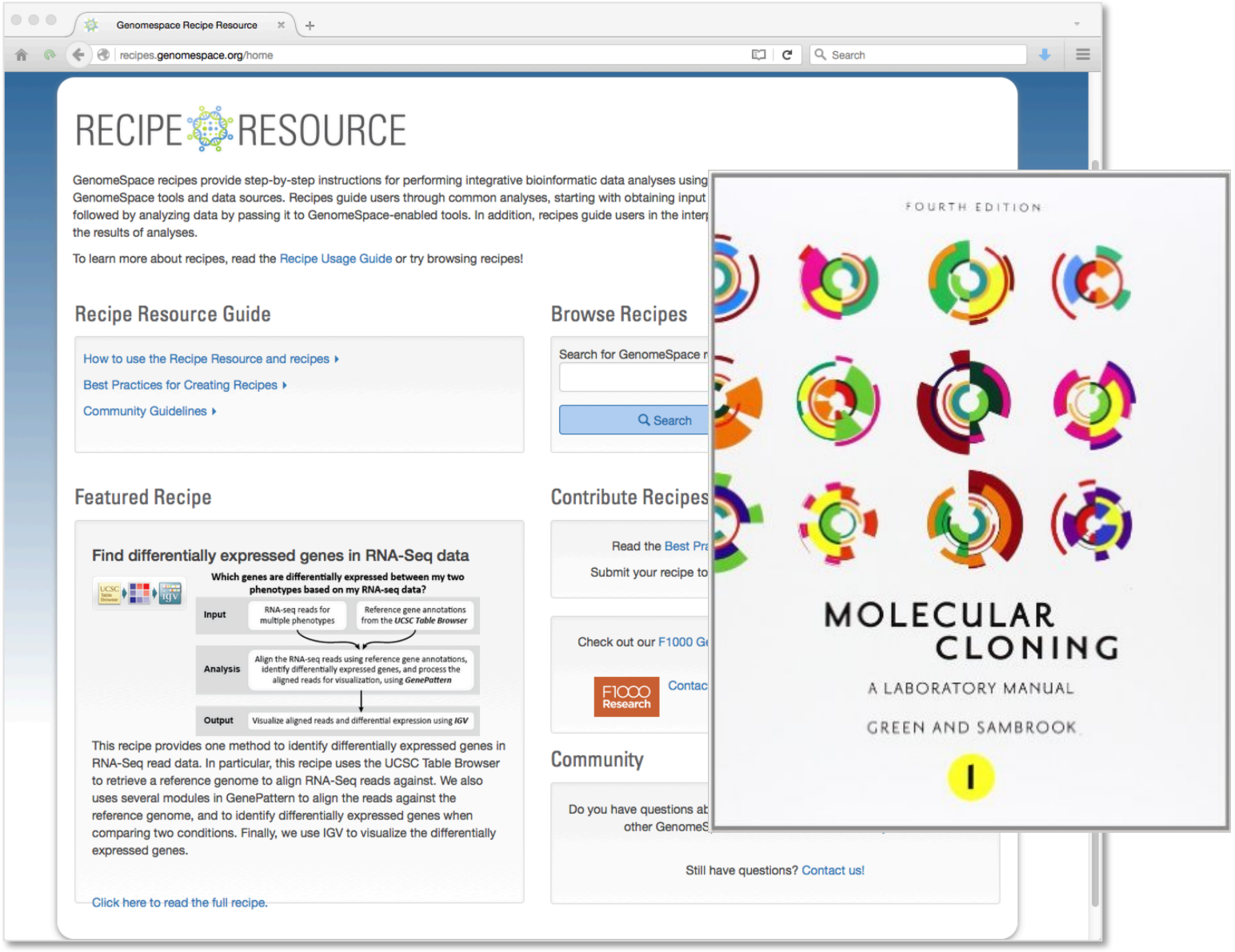


Send differentially expressed genes to Genomica



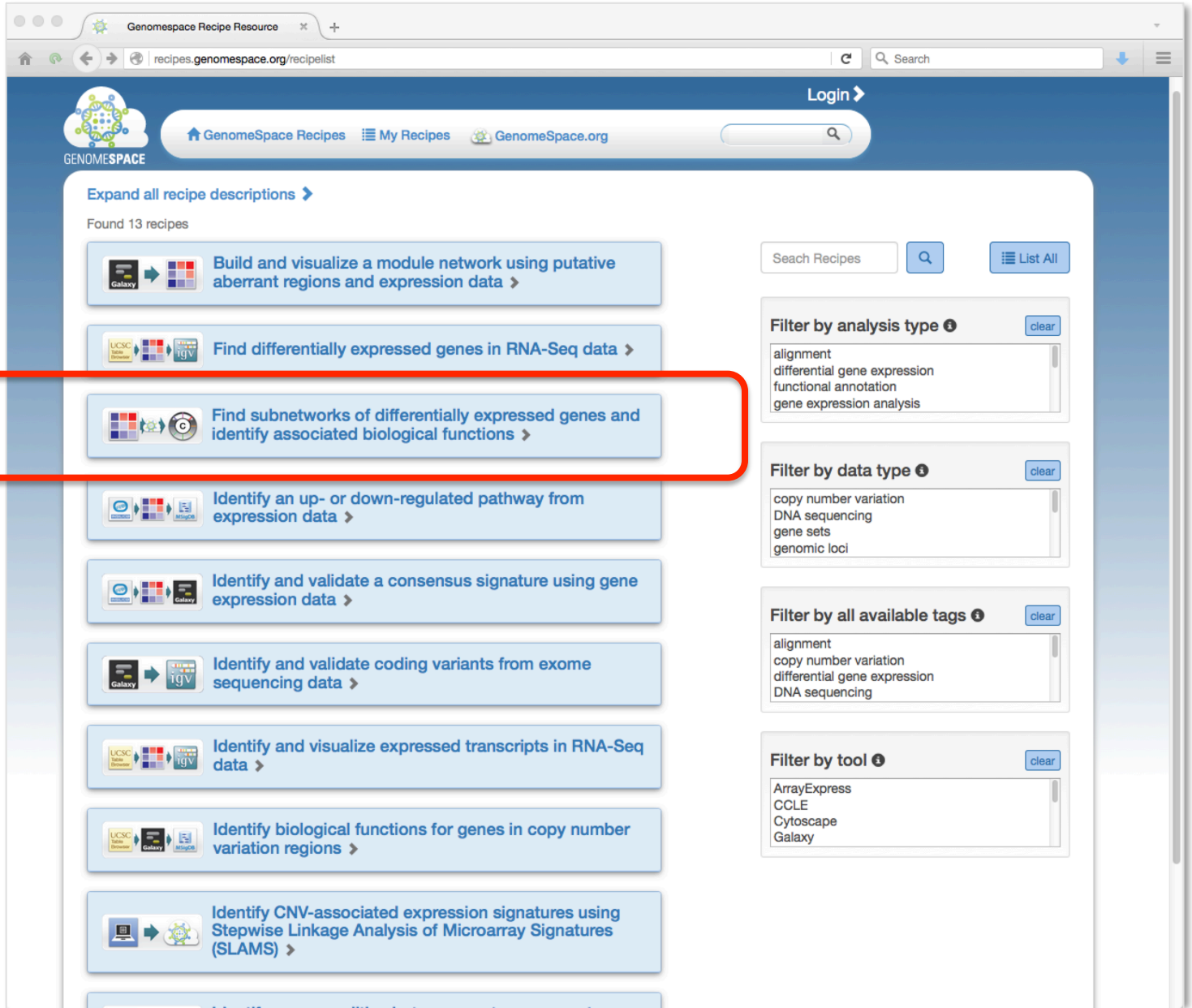

HOW DO YOU USE THIS THING?

GenomeSpace Recipe Resource

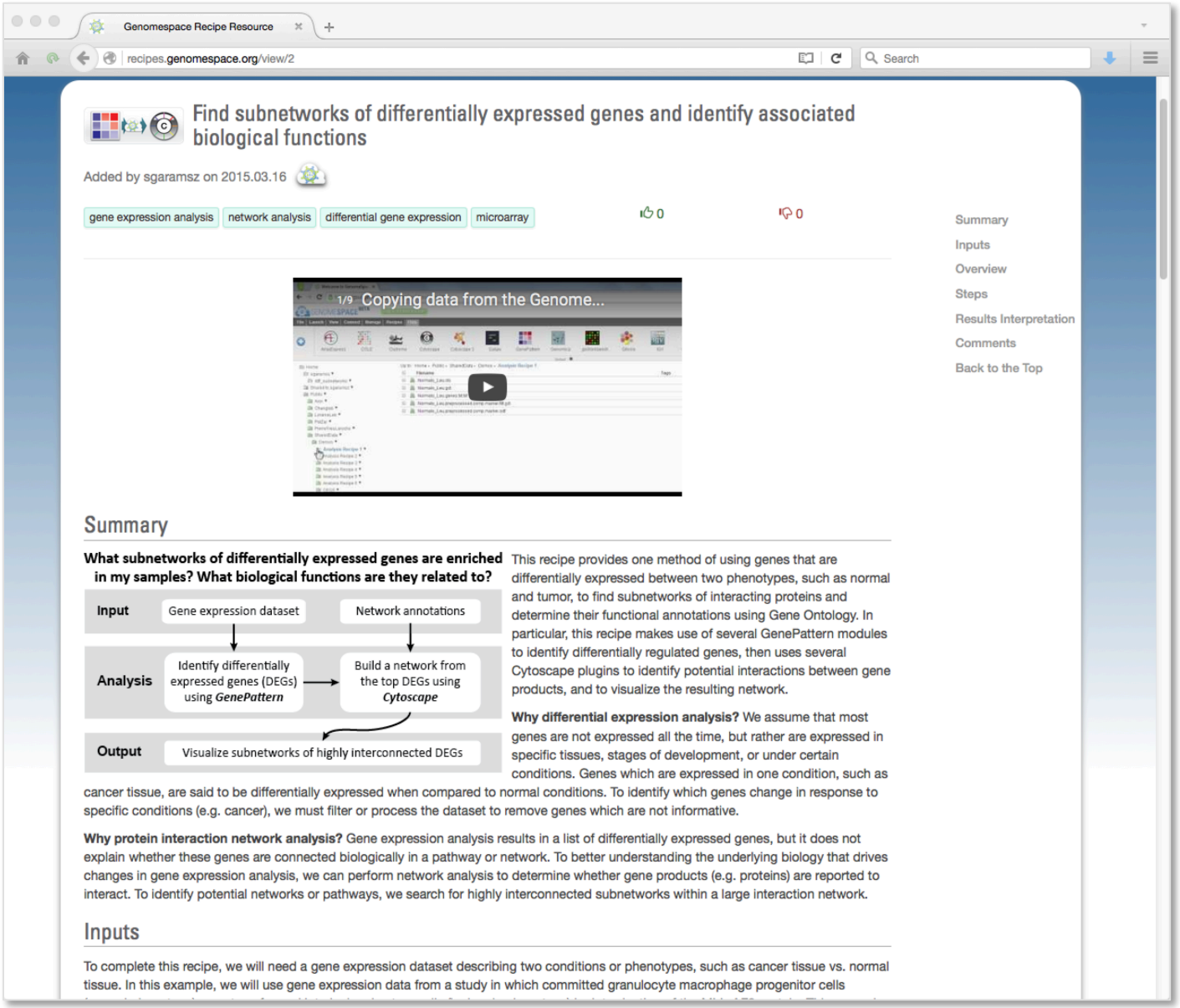


recipes.genomespace.org

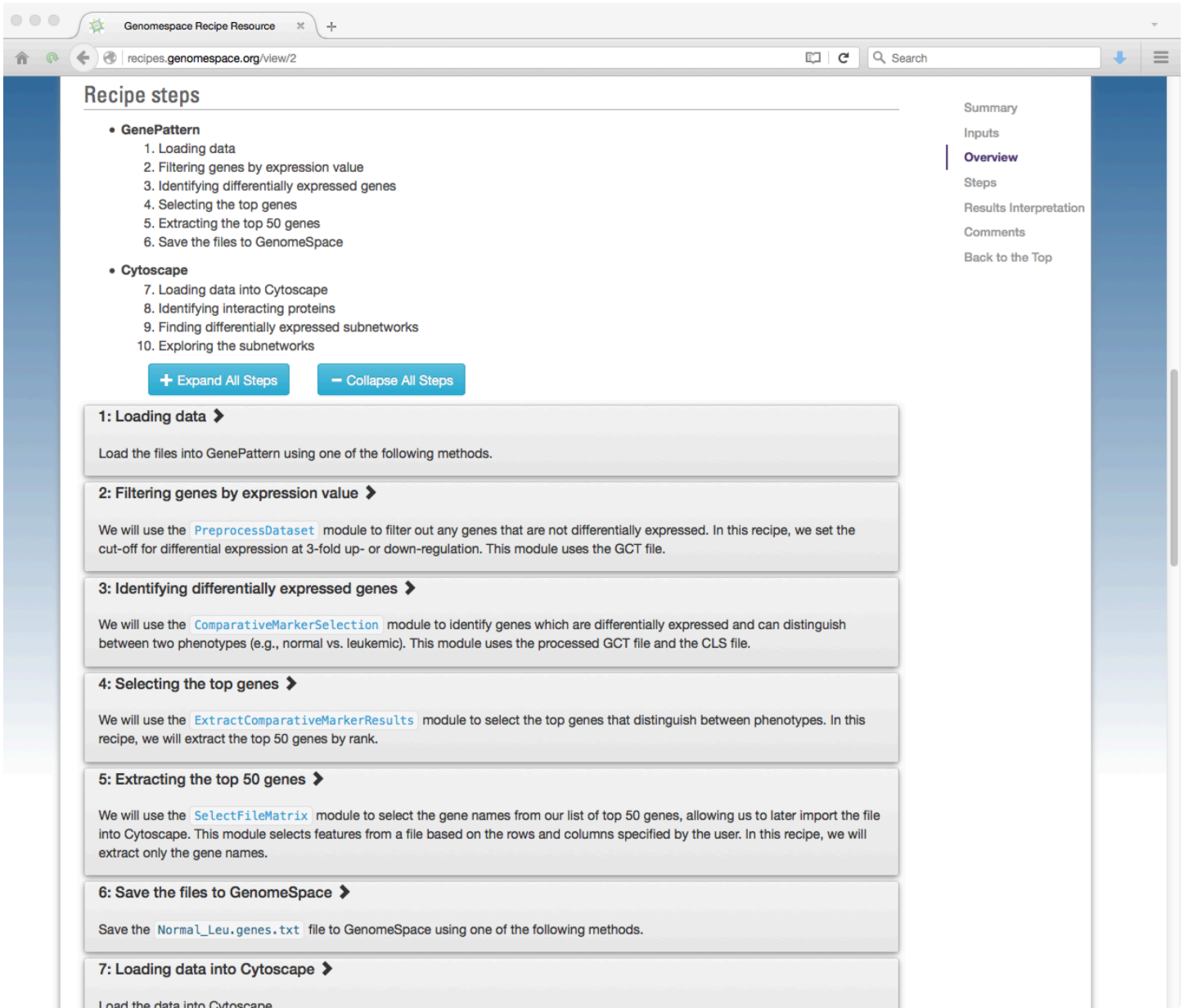
Current recipes



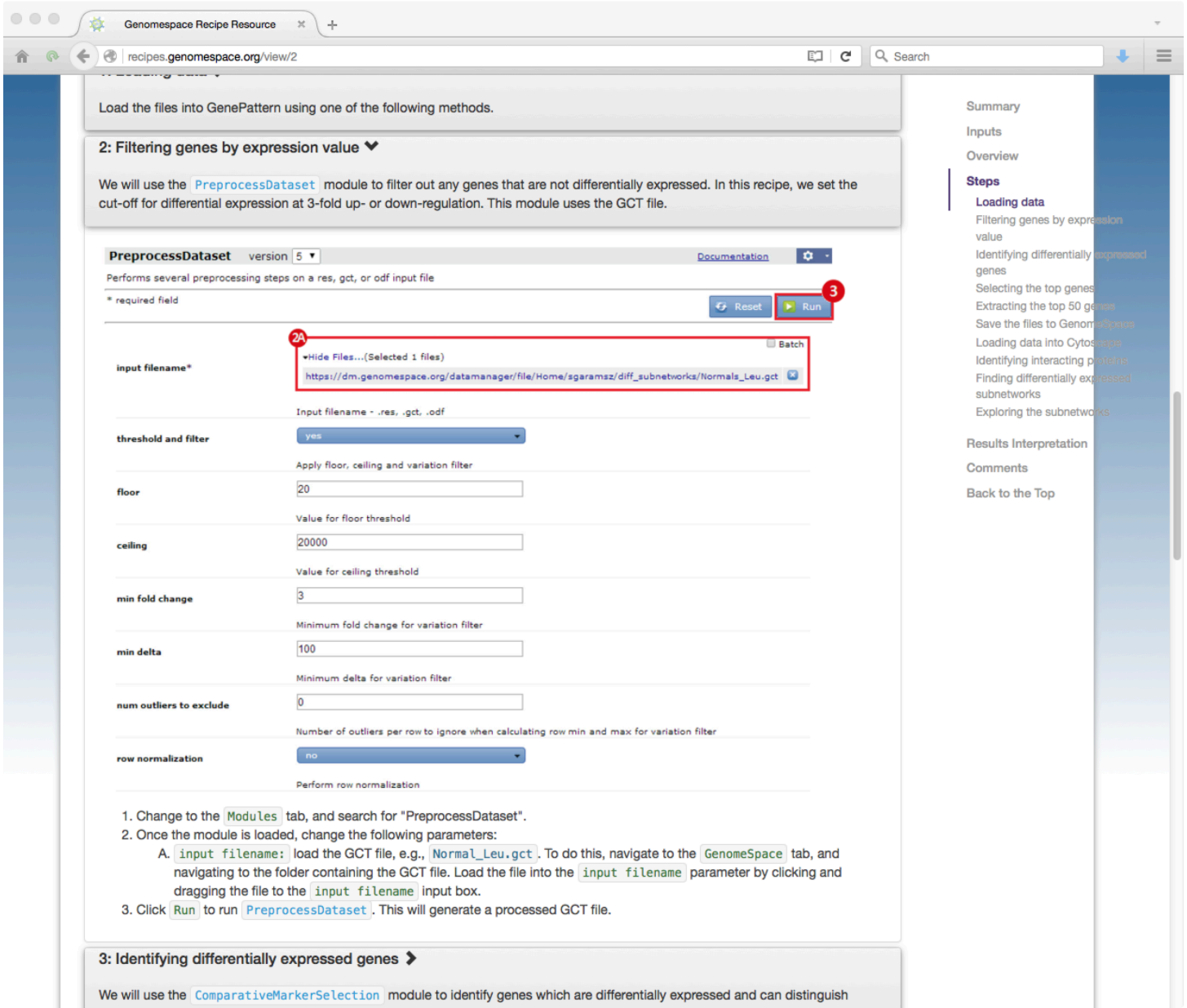
Recipe video and summary



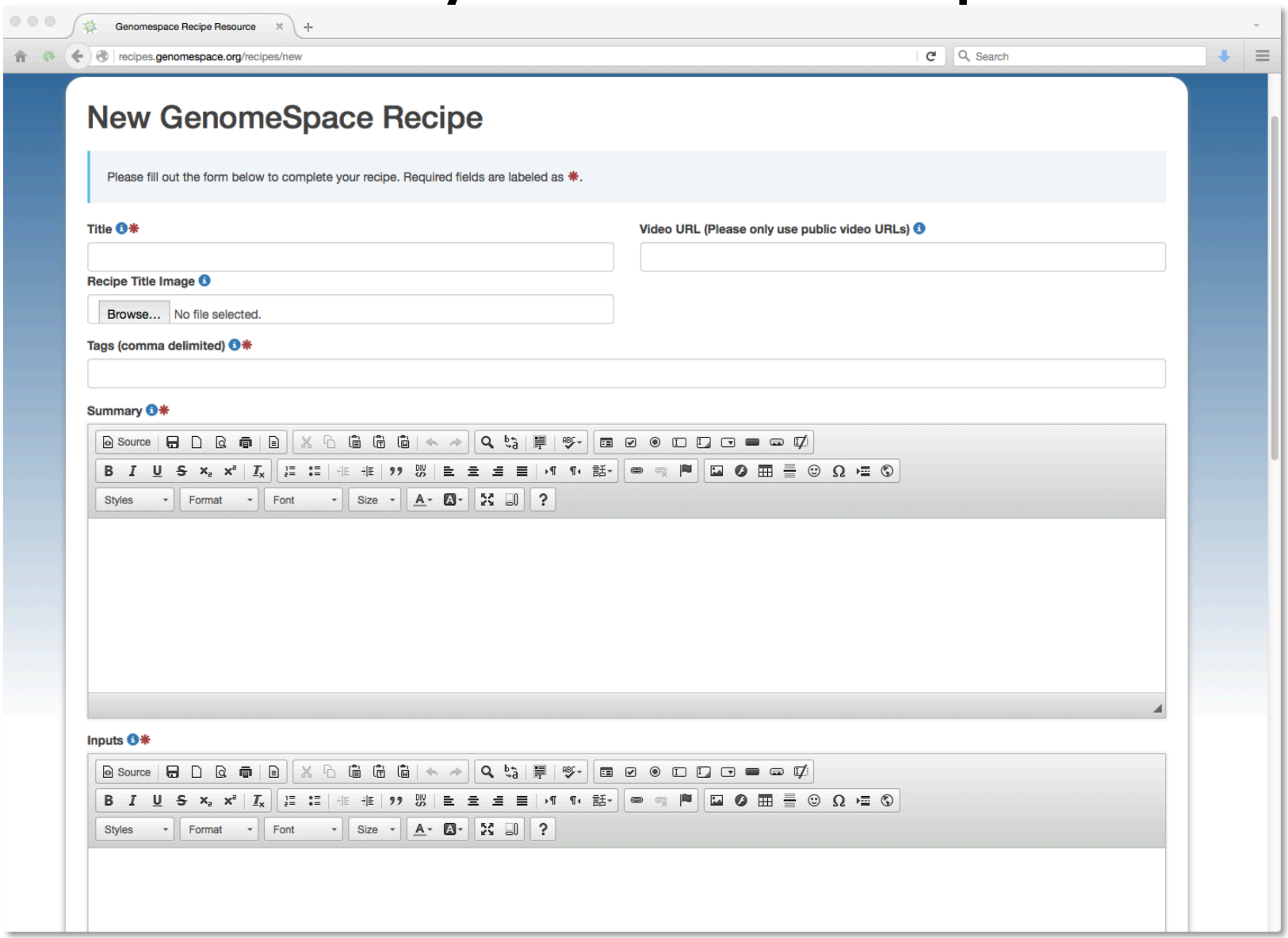
Recipe steps



Single recipe step



Create your own recipes



GenomeSpace and ITCR

- Connecting more ITCR tools
- Demo projects using GenomeSpace
- GenomeSpace as a platform for interoperable informatics across ITCR

NATIONAL CANCER INSTITUTE
Informatics Technology for Cancer Research

Acknowledgements

GenomeSpace Collaborators
Cytoscape: Trey Ideker Lab, UCSC
Galaxy: Anton Nekrutenko Lab, Penn State University
Genomica: Eran Segal Lab, Weizmann Institute
UCSC Browser Team
GenePattern Team
IGV Team

Driving Biological Projects
Howard Chang Lab – Stanford University
Aviv Regev Lab – Broad Institute

UCSD
Michael Reich
Ted Liefeld
Helga Thorvaldsdottir
Jim Robinson
Marco Ocana
Thorin Tabor
Jill Mesirov, PI

Funding

National Human Genome Research Institute

amazon web services™

GenomeSpace team: gs-dev@broadinstitute.org
www.genomespace.org

UC SANTA CRUZ UCSD PENNSTATE BROAD INSTITUTE מכון ויצמן למדע WEIZMANN INSTITUTE OF SCIENCE STANFORD UNIVERSITY National Human Genome Research Institute amazon web services™