

GENOMESPACE



(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

GenomeSpace: Home

www.genomespace.org

User Login | Admin Logout: michaelr

What is GenomeSpace? Tools Recipes Recipes (New) Documentation Developers Support About

GENOMESPACE

Frictionless connection of bioinformatics tools

Register User Login

STATUS 02.03.16 07:00AM All systems are operating normally.

Citing GenomeSpace

To cite your use of GenomeSpace, please reference Qu K, Garamszegi S, Wu F, et al. [Nature Methods](#). 2016 Jan 18. doi: 10.1038/nmeth.3732.

WHAT'S NEW

News Highlights GenomeSpace Blog

Announcing the New GenomeSpace Recipe Resource

Posted by Ted Liefeld on Tuesday, February 02, 2016 at 05:47PM

The GenomeSpace team is pleased to announce a new [GenomeSpace Recipe Resource](#) website. [More >>](#)

New Format Converters in GenomeSpace

Posted by Ted Liefeld on Friday, March 06, 2015 at 09:18AM

In response to user requests we have added the following three new format converters to the GenomeSpace ecosystem. [More >>](#)

[See All Posts](#)

@genomespace Mentions

GenomeSpace Team @genomespace 2 Feb
Announcing the new GenomeSpace Recipe Resource to support community provided recipes [bit.ly/TmatXMF](#) #bioinformatics
Expand

GenePattern @GenePattern 22 Jan
GenePattern 3.9.6 released! New read count processing modules, easier R module mgmt & more [ow.ly/XhNcg](#) #rnaseq #bioinformatics
Retweeted by GenomeSpace Team
Expand

GenomeSpace Team @genomespace 19 Jan
GenomeSpace is published in @naturemethods: [bit.ly/1UqQOlv](#) Integrative genomic analysis by interoperation of #bioinformatics tools
Expand

GenePattern 17 Dec

[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^{1,2}, Sara Garamszegi^{2,12}, Felix Wu^{2,12}, Helga Thorvaldottir², Ted Liefeld^{2,3}, Marco Ocana^{2,3}, Diego Borges-Rivera⁴, Nathalie Pochet^{2,5}, James T Robinson^{3,3}, Barry Demchak³, Tim Hull³, Gil Ben-Artzt^{2,7}, Daniel Blankenberg⁸, Galt P Barber⁹, Brian T Lee⁶, Robert M Kuhn⁶, Anton Nekrutkin⁶, Eran Segal⁶, Trey Ideker³, Michael Reich^{2,3}, Aviv Regev^{2,10}, Howard Y Chang^{1,11} & Jill P Mesirov^{2,3}

Complex biomedical analyses require the use of multiple software tools in concert and remains challenging for much of the biomedical research community. We introduce GenomeSpace ([https://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 (Cytoscape¹⁰, Galaxy⁷, GenePattern⁸, Genomica⁴, the integrative Genomics Viewer (IGV)⁹ and the UCSC Table Browser¹¹), 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^{4,8} 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 originally data transfers between tools. While original data 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



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

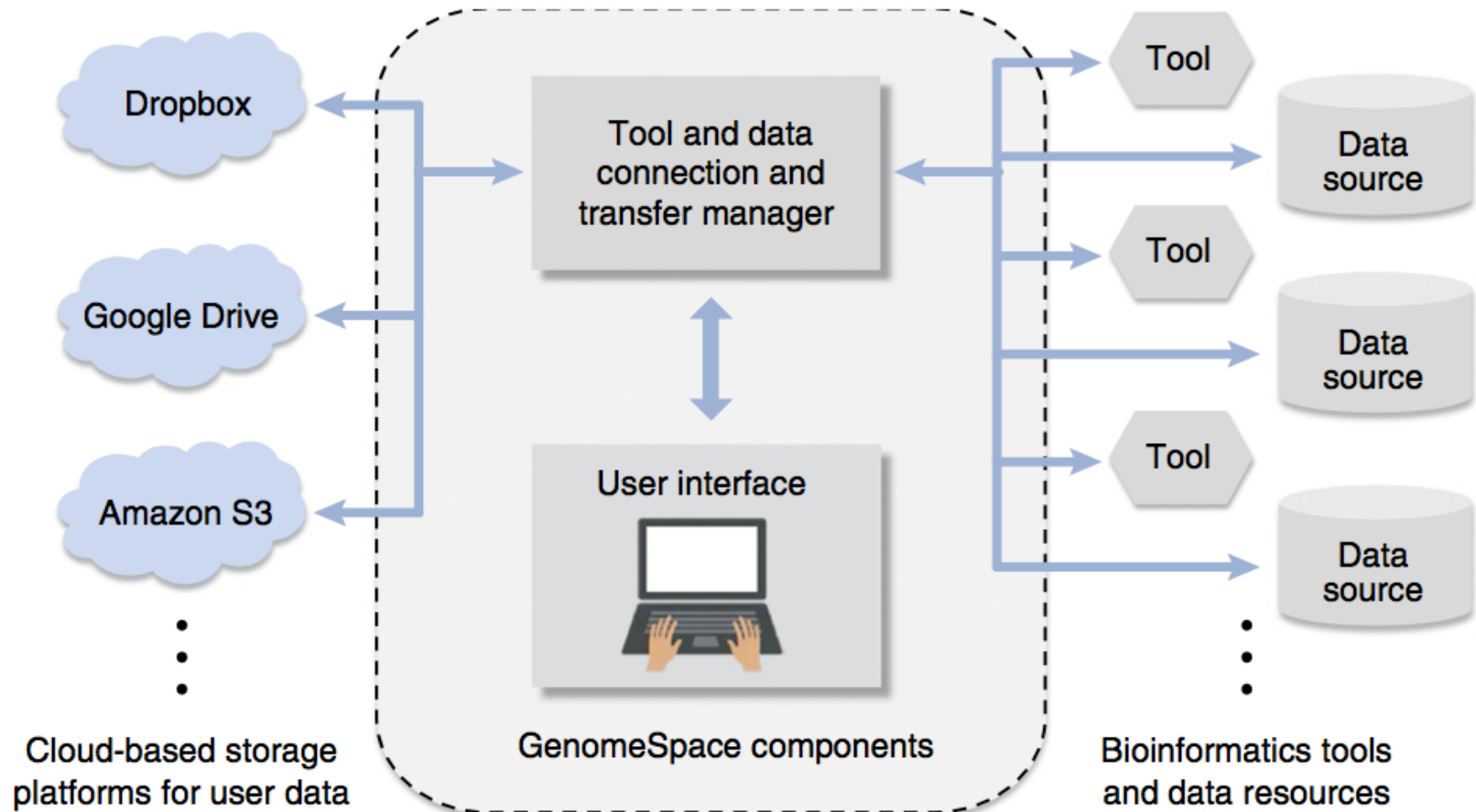


Cloud-based storage makes tools and data accessible from any location.



Tools retain their native look and feel.

GenomeSpace Architecture



GenomeSpace Tools 2/16



ArrayExpress



Cancer Cell Line Encyclopedia



Cistrome



Cytoscape



Trinity



Cytoscape 3



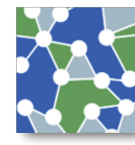
Galaxy



GenePattern



Genomica



geWorkbench



Gitools



IGV



InSilicoDB



ISACreator



Multiple Myeloma Genomics Portal



MSigDB



Project Achilles



Reactome



Synapse



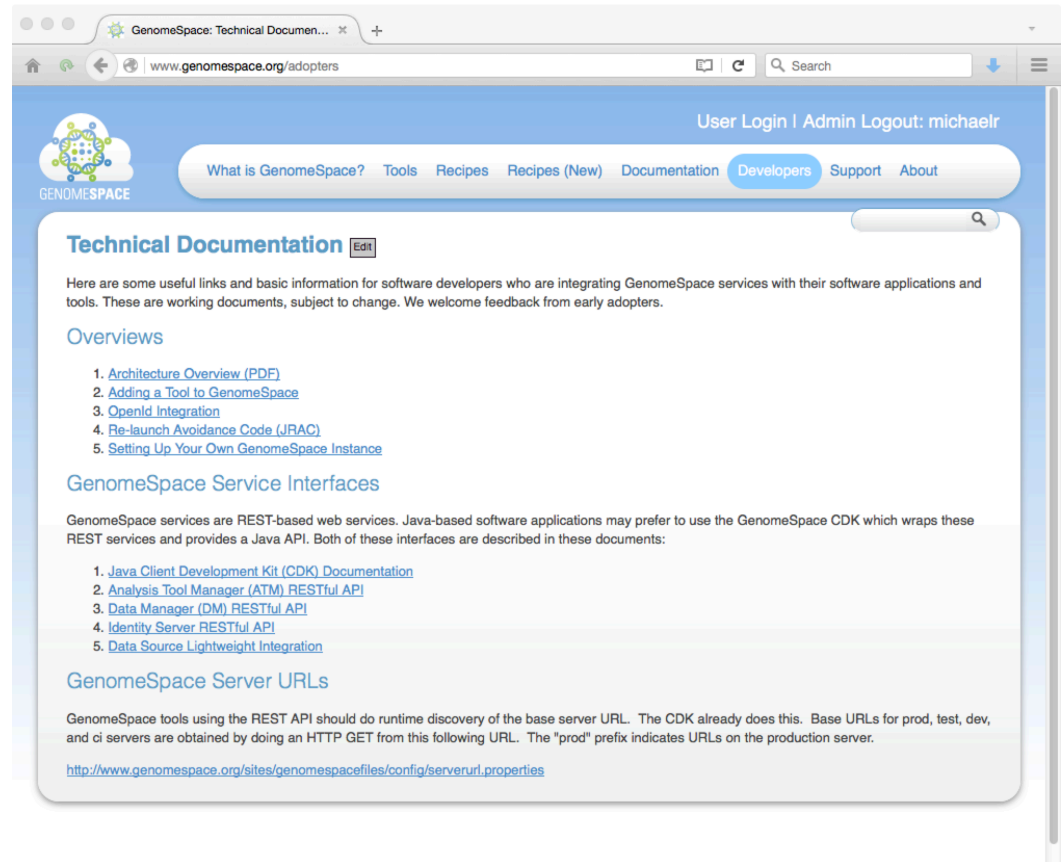
UCSC Table Browser



cBio Portal

GenomeSpace-enablement of new tools

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



The screenshot shows a web browser window displaying the GenomeSpace website. The address bar shows the URL www.genomespace.org/adopters. The page features a blue header with the GenomeSpace logo on the left and navigation links: "What is GenomeSpace?", "Tools", "Recipes", "Recipes (New)", "Documentation", "Developers" (highlighted), "Support", and "About". On the right of the header, it says "User Login | Admin Logout: michaelr". Below the header is a search bar. The main content area is titled "Technical Documentation" with an "Edit" button. The text reads: "Here are some useful links and basic information for software developers who are integrating GenomeSpace services with their software applications and tools. These are working documents, subject to change. We welcome feedback from early adopters." The page is organized into sections: "Overviews" with a list of five links (Architecture Overview (PDF), Adding a Tool to GenomeSpace, OpenId Integration, Re-launch Avoidance Code (JIRAC), and Setting Up Your Own GenomeSpace Instance); "GenomeSpace Service Interfaces" with a paragraph explaining REST-based services and a list of five links (Java Client Development Kit (CDK) Documentation, Analysis Tool Manager (ATM) RESTful API, Data Manager (DM) RESTful API, Identity Server RESTful API, and Data Source Lightweight Integration); and "GenomeSpace Server URLs" with a paragraph about runtime discovery and a link to <http://www.genomespace.org/sites/genomespacefiles/config/serverurl.properties>.

Add your own private tools

GENOMESPACE BETA

File Launch View Connect Manage

Cytoscape ArrayExpress

Home Up to:

- ted
- Shared to ted
- dropbox:Ted Liefeld
- Public
- s3:1000genomes

Edit HaploViewPlink

Name: HaploViewPlink

Description: HaploView with GenomeSpace extension

Tool Provider: Broad

Base URL: <http://vcancerportal05.broadinstitute.org:3003/HaploviewLauncher/ha>

Help URL:

File Parameter: Parameter name: **plinkFi** Required:

Allow multiple files: Multiple file Delimiter: ,

example call with dummy files
`http://vcancerportal05.broadinstitute.org:3003/HaploviewLauncher/haploview.jnlp?plinkFiles=https%3A//gs.org/File1.txt,https%3A//gs.org/File2.txt`

Available formats	Formats this tool accepts
mitab	
mut	
ndb	
nnf	
odf	
owl	

Request a new format be added to the available formats list.

Icon: `./toolicon/haploview.tiff`
 haploview.tiff

Share with groups: GS-Developers ✖

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,^{1,9} Tal Shay,¹ Garrett M. Frampton,^{2,4} Adam C.B. Drake,^{2,7} Ilya Leskov,^{2,7} Bjorn Nilsson,^{1,10} Fred Pfeffer,⁸ David Dombkowski,⁸ John W. Evans,² Ted Liefeld,¹ John S. Smutko,⁹ Jianshu Chen,^{2,7} Nir Friedman,³ Richard A. Young,^{2,4} Todd R. Golub,^{1,5,10,12} and Benjamin L. Ebert^{1,5,6,12,*}

¹Broad Institute, 7 Cambridge Center, Cambridge MA, 02142, USA

²Department of Biology, Massachusetts Institute of Technology, Cambridge MA, 02140, USA

³School of Computer Science, Hebrew University, Jerusalem 91904, Israel

⁴Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA

⁵Dana-Farber Cancer Institute, Boston, MA 02115, USA

⁶Brigham and Women's Hospital, Boston, MA 02115, USA

⁷Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139

⁸Massachusetts General Hospital, Boston, MA 02114, USA

⁹Nugen Technologies, San Carlos, CA 94070, USA

¹⁰Howard Hughes Medical Institute, Chevy Chase, MD 20815-6789, USA

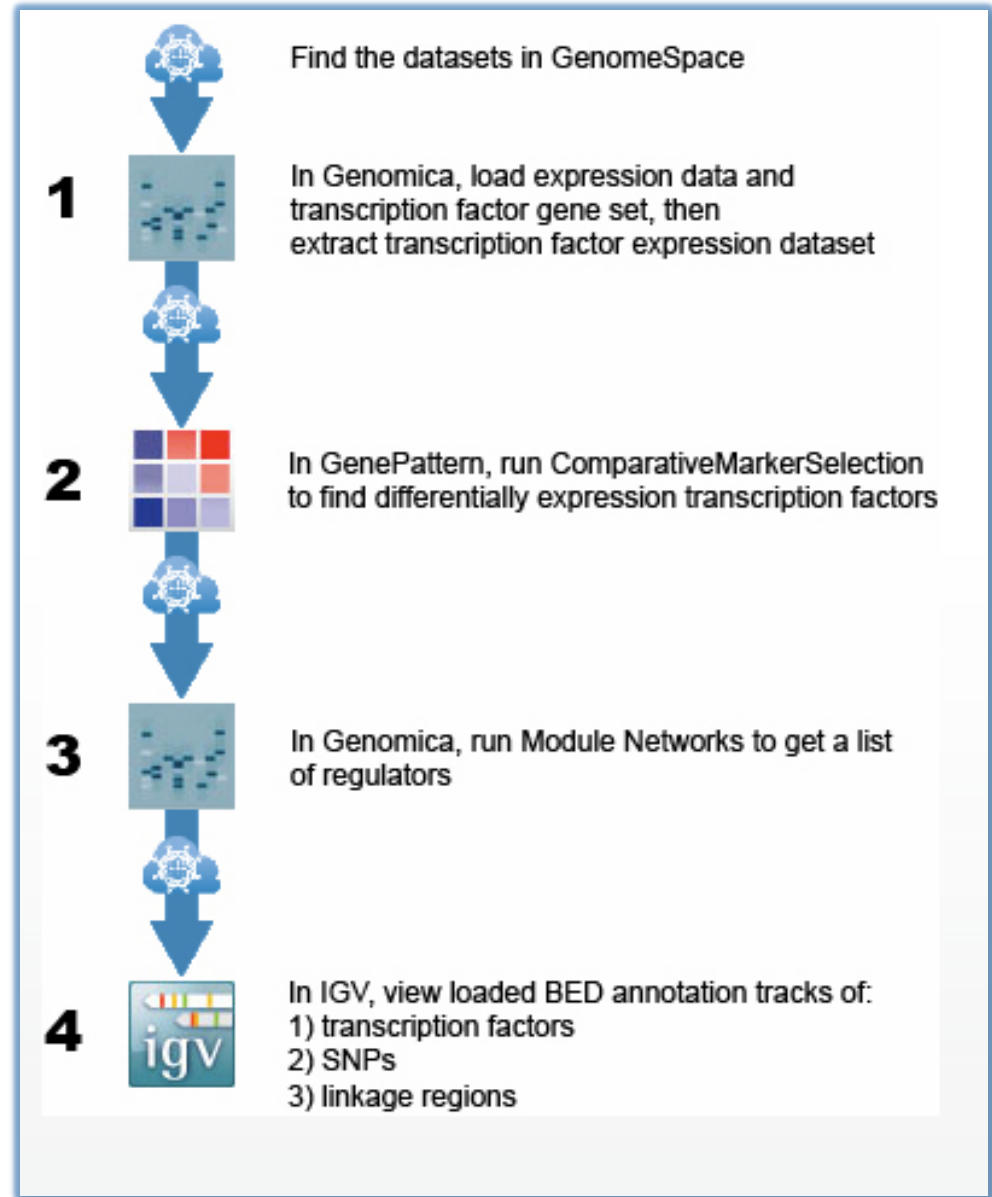
¹¹These authors contributed equally to this work

¹²These 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

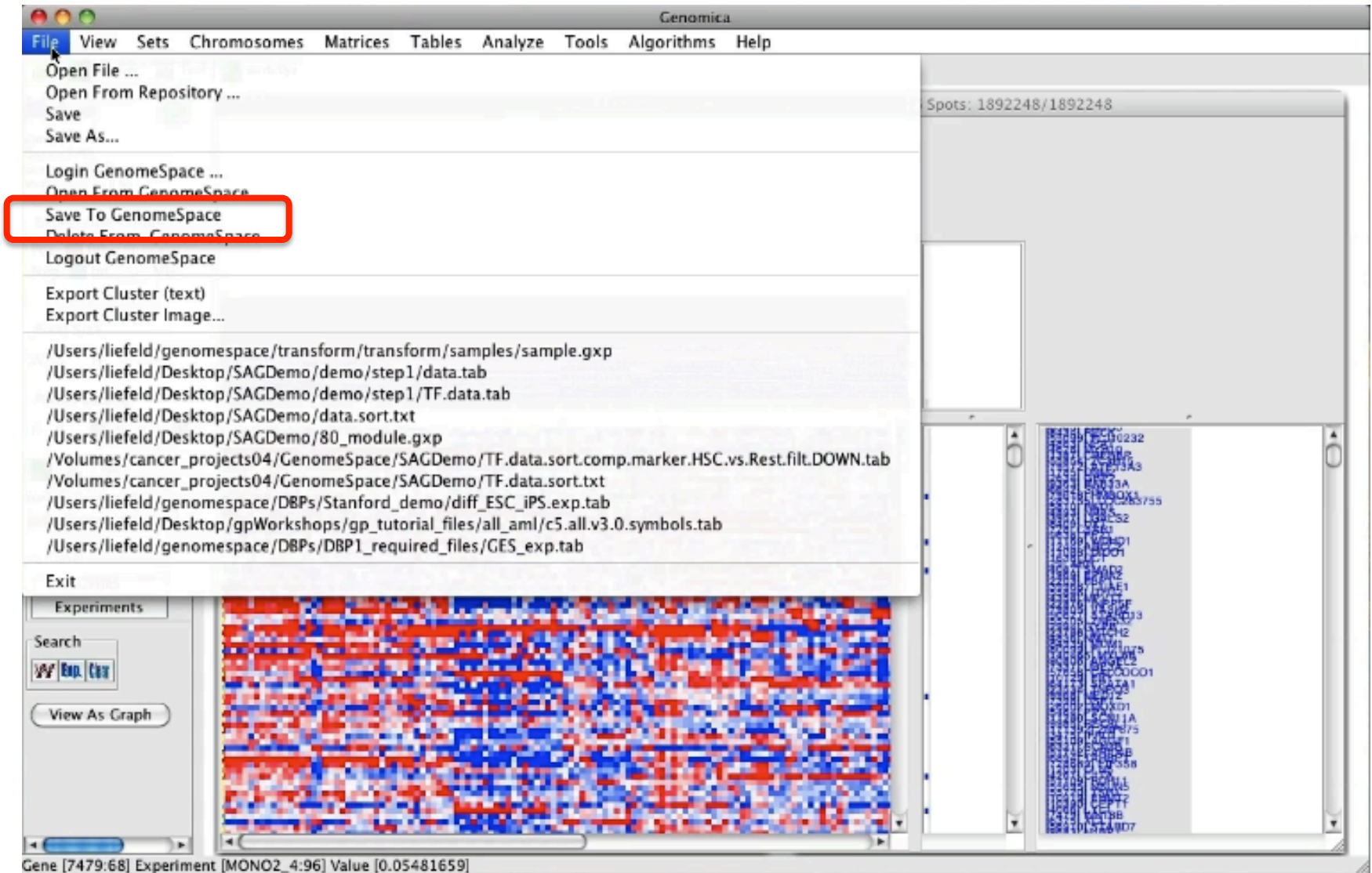


Doing this the standard way

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



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**

<input type="checkbox"/>	Filename	Tags	Owner	Size	Last Modified
<input type="checkbox"/>	DMAP.classes.cls		mmr	664 bytes	48 months ago
<input type="checkbox"/>	TF.data.comp.marker.HSC.vs.Rest.filt.gct		mmr	618 KB	48 months ago
<input type="checkbox"/>	TF.data.comp.marker.HSC.vs.Rest.filt.txt		mmr	1 KB	48 months ago
<input type="checkbox"/>	TF.data.comp.marker.HSC.vs.Rest.odf		mmr	126 KB	48 months ago
<input type="checkbox"/>	TF.data.gct		mmr	1.31 MB	48 months ago
<input type="checkbox"/>	TF.data.tab		mmr	1.31 MB	47 months ago
<input type="checkbox"/>	TF.data.txt		mmr	1.31 MB	48 months ago

File Launch View Connect Manage Recipes Help

Trinity GenePattern Genomica Cytoscape Reactome InSilicoDB Cytoscape 3 IGV ArrayExpress UCSC Table Browser Cistrome ISACreator Galaxy

Up to: Home > mmr > **step2**

Filename	Tags	Owner	Size	Last Modified
<input checked="" type="checkbox"/> DMAP.classes.cls		mmr	664 bytes	48 months ago
<input type="checkbox"/> TF.data.comp.marker.HSC.vs.Rest.filt.gct		mmr	618 KB	48 months ago
<input type="checkbox"/> TF.data.comp.marker.HSC.vs.Rest.filt.txt		mmr	1 KB	48 months ago
<input type="checkbox"/> TF.data.comp.marker.HSC.vs.Rest.odf		mmr	126 KB	48 months ago
<input type="checkbox"/> TF.data.gct		mmr	1.31 MB	48 months ago
<input checked="" type="checkbox"/> TF.data.tab		mmr	1.31 MB	47 months ago
<input type="checkbox"/> TF.data.txt		mmr	1.31 MB	48 months ago

File Launch View Connect Manage Recipes Help

Trinity GenePattern Genomica Cytoscape Reactome InSilicoDB Cytoscape 3 IGV ArrayExpress UCSC Table Browser Cistrome ISACreator Galaxy

- 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
<input checked="" type="checkbox"/> DMAP.classes.cls		mmr	664 bytes	48 months ago
<input type="checkbox"/> TF.data.comp.marker.HSC.vs.Rest.filt.gct		mmr	618 KB	48 months ago
<input type="checkbox"/> TF.data.comp.marker.HSC.vs.Rest.filt.gct		mmr	1 KB	48 months ago
<input type="checkbox"/> TF.data.comp.marker.HSC.vs.Rest.filt.gct		mmr	126 KB	48 months ago
<input type="checkbox"/> TF.data.gct		mmr	1.31 MB	48 months ago
<input checked="" type="checkbox"/> TF.data.tab		mmr	1.31 MB	47 months ago
<input type="checkbox"/> TF.data.txt		mmr	1.31 MB	48 months ago

Launch GenePattern

Currently Selected Files

- /mmr/step2/DMAP.classes.cls
- /mmr/step2/TF.data.tab

Drop files here

TF.data.tab will be converted to an acceptable format.

Perform differential expression analysis in GenePattern

Welcome to GenomeSpace | Receive GenomeSpace File | Search | mmrtest2 | 0.0 KB / 30 GB

GenePattern

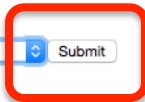
Modules & Pipelines | Suites | Job Results | Resources | Help | GenomeSpace

GenomeSpace has sent you a file

DMAP.classes.cls
Save a Copy: Upload Directory Submit

TF.data.tab *Converted to gct*
Save a Copy: Upload Directory Submit

Send to Module: ComparativeMarkerSelection Submit



Perform differential expression analysis in GenePattern

Welcome to GenomeSpace

GenePattern - ComparativeMarkerS...

genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=urn:lsid:broad.mit.edu:cancer.software.genepattern.module.analysis:00044:10

GenePattern

mmrtest2

0.0 KB / 30 GB

Modules & Pipelines Suites Job Results Resources Help GenomeSpace

Modules Jobs Files GenomeSpace

Search Modules & Pipelines

No Jobs Processing Browse Modules

Favorite Modules

Drag Modules Here

Recent Modules

No Recent Modules

ComparativeMarkerSelection version 10 Documentation

Identify differentially expressed genes that can discriminate between distinct classes of samples.

* required field

Reset Run

input file*

Hide Files...(Selected 1 files)

https://dm.genomespace.org/datamanager/file/Home/mmr/step2/TF.data.tab?dataformat=http://www.genomespace.org/datamanager/dataformat/gct

The input file - .res, .gct. Note that if your data is log transformed, you will need to set the "log transformed data" parameter below to "yes"

cls file*

Hide Files...(Selected 1 files)

https://dm.genomespace.org/datamanager/file/Home/mmr/step2/DMAP.classes.cls

The class file - .cls

confounding variable cls file

Upload File... Add Path or URL... Drag Files Here

2GB file upload limit using the Upload File... button. For files > 2GB upload from the Files tab.

The class file containing the confounding variable.

test direction* 2 Sided

The test to perform

test statistic* T-Test

The statistic to use

min_std

About GenePattern | Contact Us

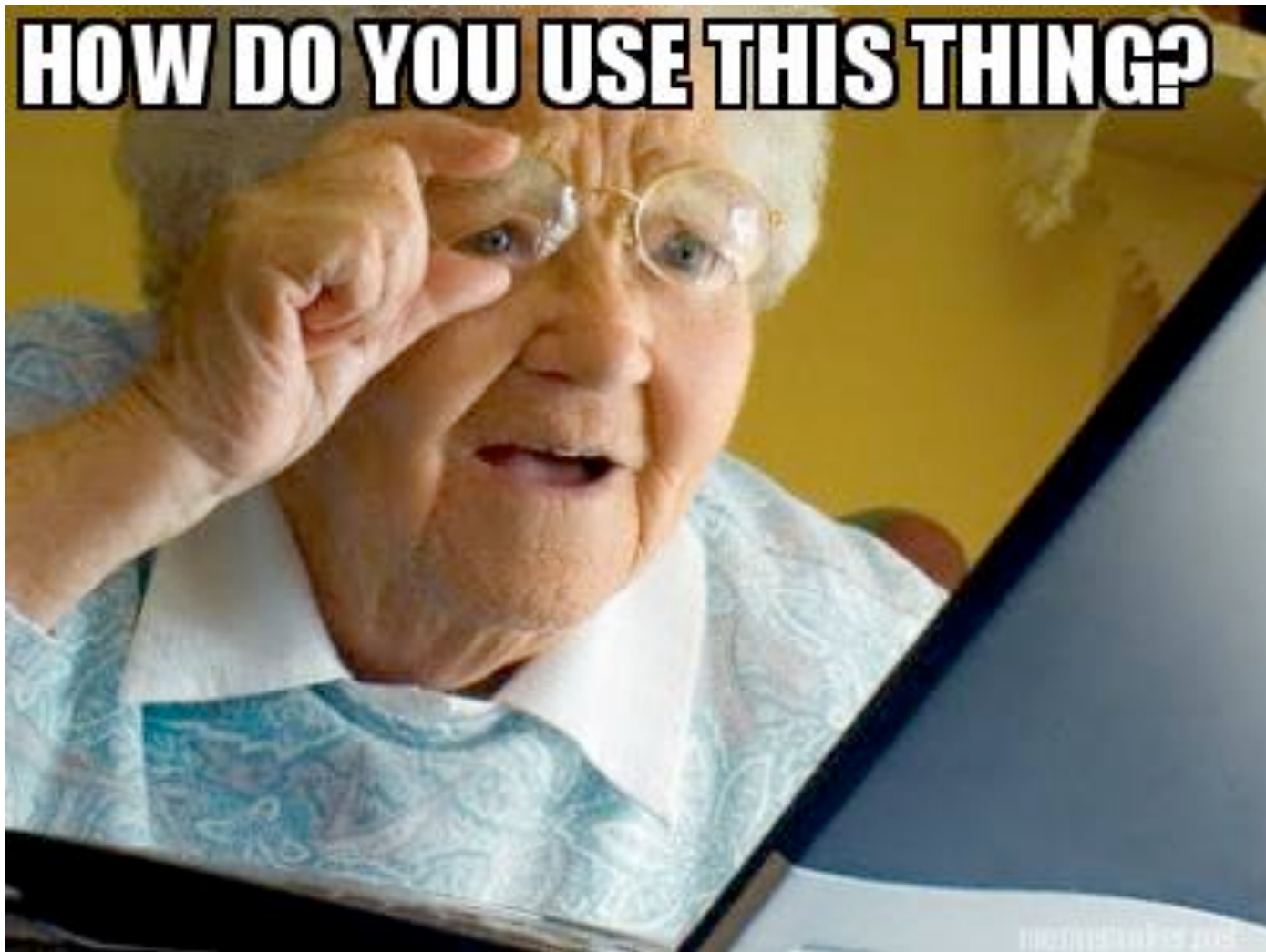
©2003-2015 Broad Institute, MIT

Send differentially expressed genes to Genomica

The screenshot shows the GenePattern web interface. On the left, a file tree is visible with the file 'TF.data.gct' selected and highlighted with a red box. The main panel displays the 'TF.data.gct' file details, including options to 'Delete File', 'Save File', and 'Send to Genomica'. The 'Send to Genomica' option is highlighted with a red box. Below this, there are sections for 'Send to GenomeSpace Tool' and 'Send to Module as tab'.

The right side of the screenshot shows a list of modules and their descriptions. The 'Send to Genomica' module is highlighted with a red box. Below it, there are sections for 'Send to Galaxy', 'Send to Gitools', 'Send to IGV', and 'Send to Trinity'. At the bottom, there are sections for 'Send to Module as tab' and 'ChIPSeq.CreateHeatmap' and 'ChIPSeq.CreatePlots'.

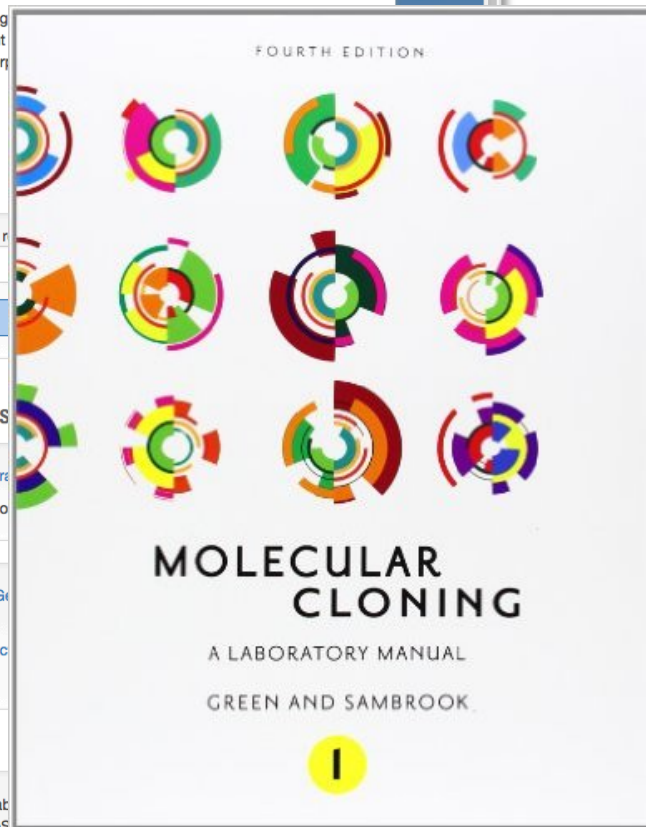
HOW DO YOU USE THIS THING?



GenomeSpace Recipe Resource

The screenshot shows the homepage of the GenomeSpace Recipe Resource. The browser address bar displays `recipes.genomespace.org/home`. The page features a navigation menu, a search bar, and several content sections:

- RECIPE RESOURCE**: The main header with a logo.
- Introduction**: A paragraph explaining that GenomeSpace recipes provide step-by-step instructions for integrative bioinformatic data analyses.
- Recipe Resource Guide**: A sidebar with links for "How to use the Recipe Resource and recipes", "Best Practices for Creating Recipes", and "Community Guidelines".
- Browse Recipes**: A search box with a "Search" button.
- Featured Recipe**: A detailed section for "Find differentially expressed genes in RNA-Seq data". It includes a flowchart with the following steps:
 - Input**: RNA-seq reads for multiple phenotypes and Reference gene annotations from the UCSC Table Browser.
 - Analysis**: Align the RNA-seq reads using reference gene annotations, identify differentially expressed genes, and process the aligned reads for visualization, using GenePattern.
 - Output**: Visualize aligned reads and differential expression using IGV.
- Contribute Recipes**: A section with a "Submit your recipe to" button and a link to "Read the Best Practices".
- Community**: A section with a "Contact us" button and a link to "Check out our F1000 Genomics article".



recipes.genomespace.org

Current recipes

The screenshot shows the Genomespace Recipe Resource website. The browser address bar displays `recipes.genomespace.org/recipeList`. The page header includes the Genomespace logo, navigation links for 'GenomeSpace Recipes', 'My Recipes', and 'GenomeSpace.org', and a 'Login' button. Below the header, there is a search bar and a 'List All' button. The main content area is titled 'Expand all recipe descriptions' and 'Found 13 recipes'. A list of 13 recipe cards is displayed, each with an icon and a title. The third recipe card, 'Find subnetworks of differentially expressed genes and identify associated biological functions', is highlighted with a red rectangular box. To the right of the recipe list, there are four filter sections: 'Filter by analysis type', 'Filter by data type', 'Filter by all available tags', and 'Filter by tool'. Each filter section has a 'clear' button and a scrollable list of options.

Genomespace Recipe Resource

recipes.genomespace.org/recipeList

GenomeSpace Recipes My Recipes GenomeSpace.org

Expand all recipe descriptions

Found 13 recipes

Build and visualize a module network using putative aberrant regions and expression data

Find differentially expressed genes in RNA-Seq data

Find subnetworks of differentially expressed genes and identify associated biological functions

Identify an up- or down-regulated pathway from expression data

Identify and validate a consensus signature using gene expression data

Identify and validate coding variants from exome sequencing data

Identify and visualize expressed transcripts in RNA-Seq data

Identify biological functions for genes in copy number variation regions

Identify CNV-associated expression signatures using Stepwise Linkage Analysis of Microarray Signatures (SLAMS)

Search Recipes

List All

Filter by analysis type

- alignment
- differential gene expression
- functional annotation
- gene expression analysis

Filter by data type

- copy number variation
- DNA sequencing
- gene sets
- genomic loci

Filter by all available tags

- alignment
- copy number variation
- differential gene expression
- DNA sequencing

Filter by tool

- ArrayExpress
- CCLC
- Cytoscape
- Galaxy

Recipe video and summary

The screenshot shows a web browser window with the address bar at `recipes.genomespace.org/view/2`. The page title is "Find subnetworks of differentially expressed genes and identify associated biological functions". It was added by `sgaramsz` on 2015.03.16. The recipe is categorized under "gene expression analysis", "network analysis", "differential gene expression", and "microarray". There are 0 likes and 0 dislikes. A video player is embedded in the center, showing a thumbnail for "1/9 - Copying data from the Genome...". On the right side, there is a table of contents with links to "Summary", "Inputs", "Overview", "Steps", "Results Interpretation", "Comments", and "Back to the Top".

Summary

What subnetworks of differentially expressed genes are enriched in my samples? What biological functions are they related to?

Input

- Gene expression dataset
- Network annotations

Analysis

- Identify differentially expressed genes (DEGs) using *GenePattern*
- Build a network from the top DEGs using *Cytoscape*

Output

- Visualize subnetworks of highly interconnected DEGs

Why differential expression analysis? We assume that most genes are not expressed all the time, but rather are expressed in specific tissues, stages of development, or under certain conditions. Genes which are expressed in one condition, such as cancer tissue, are said to be differentially expressed when compared to normal conditions. To identify which genes change in response to specific conditions (e.g. cancer), we must filter or process the dataset to remove genes which are not informative.

Why protein interaction network analysis? Gene expression analysis results in a list of differentially expressed genes, but it does not explain whether these genes are connected biologically in a pathway or network. To better understanding the underlying biology that drives changes in gene expression analysis, we can perform network analysis to determine whether gene products (e.g. proteins) are reported to interact. To identify potential networks or pathways, we search for highly interconnected subnetworks within a large interaction network.

Inputs

To complete this recipe, we will need a gene expression dataset describing two conditions or phenotypes, such as cancer tissue vs. normal tissue. In this example, we will use gene expression data from a study in which committed granulocyte macrophage progenitor cells

Recipe steps

The screenshot shows a web browser window with the URL `recipes.genomespace.org/view/2`. The page title is "Recipe steps". On the right side, there is a navigation menu with the following items: Summary, Inputs, Overview (highlighted), Steps, Results Interpretation, Comments, and Back to the Top.

The main content area lists 10 steps, grouped into two categories:

- GenePattern**
 1. Loading data
 2. Filtering genes by expression value
 3. Identifying differentially expressed genes
 4. Selecting the top genes
 5. Extracting the top 50 genes
 6. Save the files to GenomeSpace
- Cytoscape**
 7. Loading data into Cytoscape
 8. Identifying interacting proteins
 9. Finding differentially expressed subnetworks
 10. Exploring the subnetworks

Below the list are two buttons: "+ Expand All Steps" and "- Collapse All Steps".

The first six steps are expanded and show detailed instructions:

- 1: Loading data** ▶
Load the files into GenePattern using one of the following methods.
- 2: Filtering genes by expression value** ▶
We will use the [PreprocessDataset](#) module to filter out any genes that are not differentially expressed. In this recipe, we set the cut-off for differential expression at 3-fold up- or down-regulation. This module uses the GCT file.
- 3: Identifying differentially expressed genes** ▶
We will use the [ComparativeMarkerSelection](#) module to identify genes which are differentially expressed and can distinguish between two phenotypes (e.g., normal vs. leukemic). This module uses the processed GCT file and the CLS file.
- 4: Selecting the top genes** ▶
We will use the [ExtractComparativeMarkerResults](#) module to select the top genes that distinguish between phenotypes. In this recipe, we will extract the top 50 genes by rank.
- 5: Extracting the top 50 genes** ▶
We will use the [SelectFileMatrix](#) module to select the gene names from our list of top 50 genes, allowing us to later import the file into Cytoscape. This module selects features from a file based on the rows and columns specified by the user. In this recipe, we will extract only the gene names.
- 6: Save the files to GenomeSpace** ▶
Save the [Norma_Leu.genes.txt](#) file to GenomeSpace using one of the following methods.

Step 7 is partially visible:

- 7: Loading data into Cytoscape** ▶
Load the data into Cytoscape

Single recipe step

The screenshot shows a web browser window with the URL `recipes.genomespace.org/view/2`. The page title is "Genomespace Recipe Resource". The main content area is titled "2: Filtering genes by expression value" and contains instructions for using the "PreprocessDataset" module. The module configuration form includes fields for "input filename", "threshold and filter", "floor", "ceiling", "min fold change", "min delta", "num outliers to exclude", and "row normalization". The "input filename" field is highlighted with a red box and labeled "2A", and the "Run" button is labeled "3". A sidebar on the right contains a "Steps" section with a list of tasks: "Loading data", "Filtering genes by expression value", "Identifying differentially expressed genes", "Selecting the top genes", "Extracting the top 50 genes", "Save the files to Genomespace", "Loading data into Cytoscape", "Identifying interacting proteins", "Finding differentially expressed subnetworks", and "Exploring the subnetworks". Below the "Steps" section are "Results Interpretation", "Comments", and "Back to the Top".

Load the files into GenePattern using one of the following methods.

2: Filtering genes by expression value

We will use the `PreprocessDataset` module to filter out any genes that are not differentially expressed. In this recipe, we set the cut-off for differential expression at 3-fold up- or down-regulation. This module uses the GCT file.

PreprocessDataset version 5 [Documentation](#)

Performs several preprocessing steps on a res, gct, or odf input file

* required field Reset Run

input filename* Hide Files... (Selected 1 files) Batch
`https://dm.genomespace.org/datamanager/file/Home/sgaramsz/diff_subnetworks/Normals_Leu.gct`

Input filename - .res, .gct, .odf

threshold and filter yes

Apply floor, ceiling and variation filter

floor
Value for floor threshold

ceiling
Value for ceiling threshold

min fold change
Minimum fold change for variation filter

min delta
Minimum delta for variation filter

num outliers to exclude
Number of outliers per row to ignore when calculating row min and max for variation filter

row normalization no
Perform row normalization

1. Change to the `Modules` tab, and search for "PreprocessDataset".
2. Once the module is loaded, change the following parameters:
 - A. `input filename`: load the GCT file, e.g., `Normal_Leu.gct`. To do this, navigate to the `GenomeSpace` tab, and navigating to the folder containing the GCT file. Load the file into the `input filename` parameter by clicking and dragging the file to the `input filename` input box.
3. Click `Run` to run `PreprocessDataset`. This will generate a processed GCT file.

3: Identifying differentially expressed genes

We will use the `ComparativeMarkerSelection` module to identify genes which are differentially expressed and can distinguish

Summary
Inputs
Overview
Steps
Loading data
Filtering genes by expression value
Identifying differentially expressed genes
Selecting the top genes
Extracting the top 50 genes
Save the files to Genomespace
Loading data into Cytoscape
Identifying interacting proteins
Finding differentially expressed subnetworks
Exploring the subnetworks
Results Interpretation
Comments
Back to the Top

Create your own recipes

Genomespace Recipe Resource

recipes.genomespace.org/recipes/new

New GenomeSpace Recipe

Please fill out the form below to complete your recipe. Required fields are labeled as *.

Title *

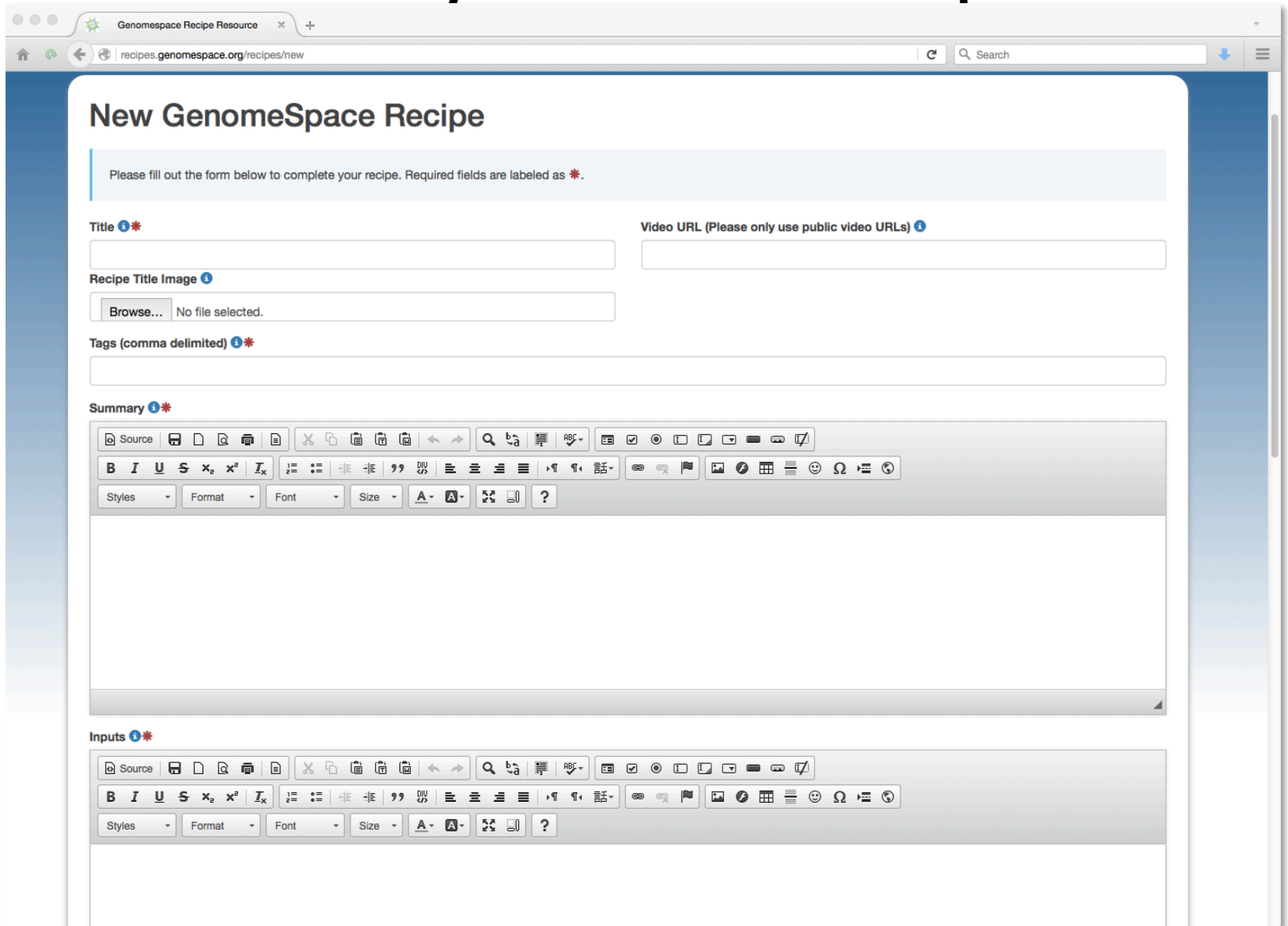
Video URL (Please only use public video URLs)

Recipe Title Image No file selected.

Tags (comma delimited)

Summary *

Inputs *



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