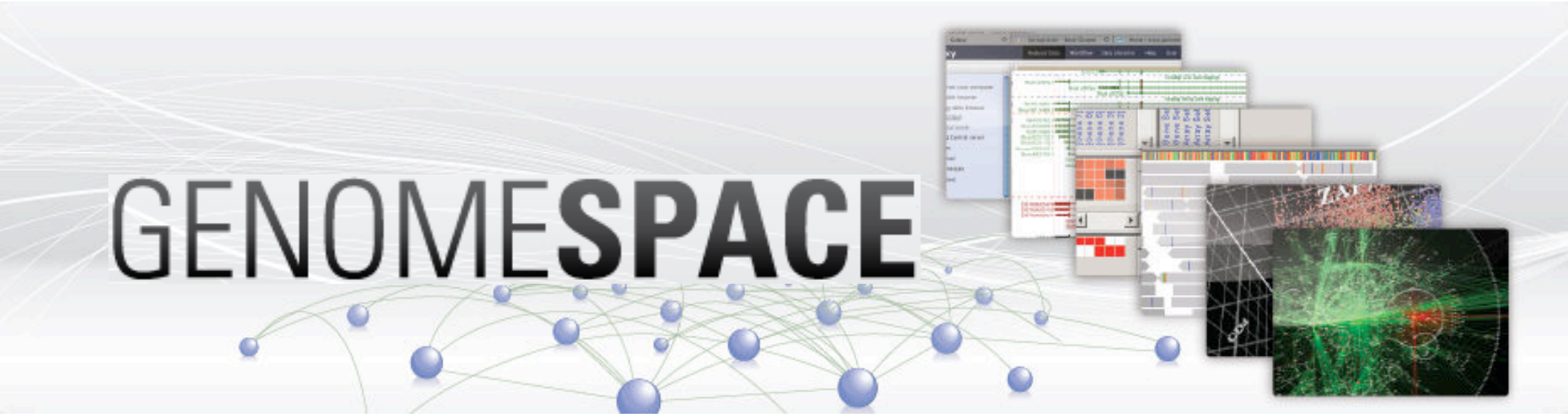


(tools x connections) + recipes = integrative analysis

Ted Liefeld

Broad Institute

July 15, 2014

How many bioinformatics tools and databases are there?

~10,000 tools
~ 5,000 databases

What could researchers accomplish if those tools could talk to each other?

Case study:
Studying the regulatory control of human hematopoiesis

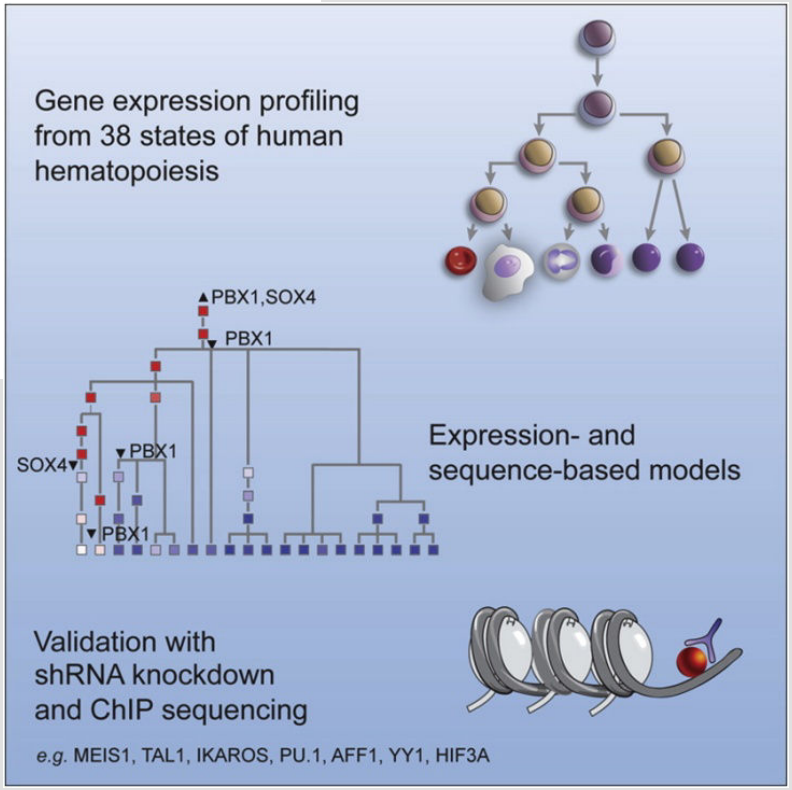
Cell

Resource

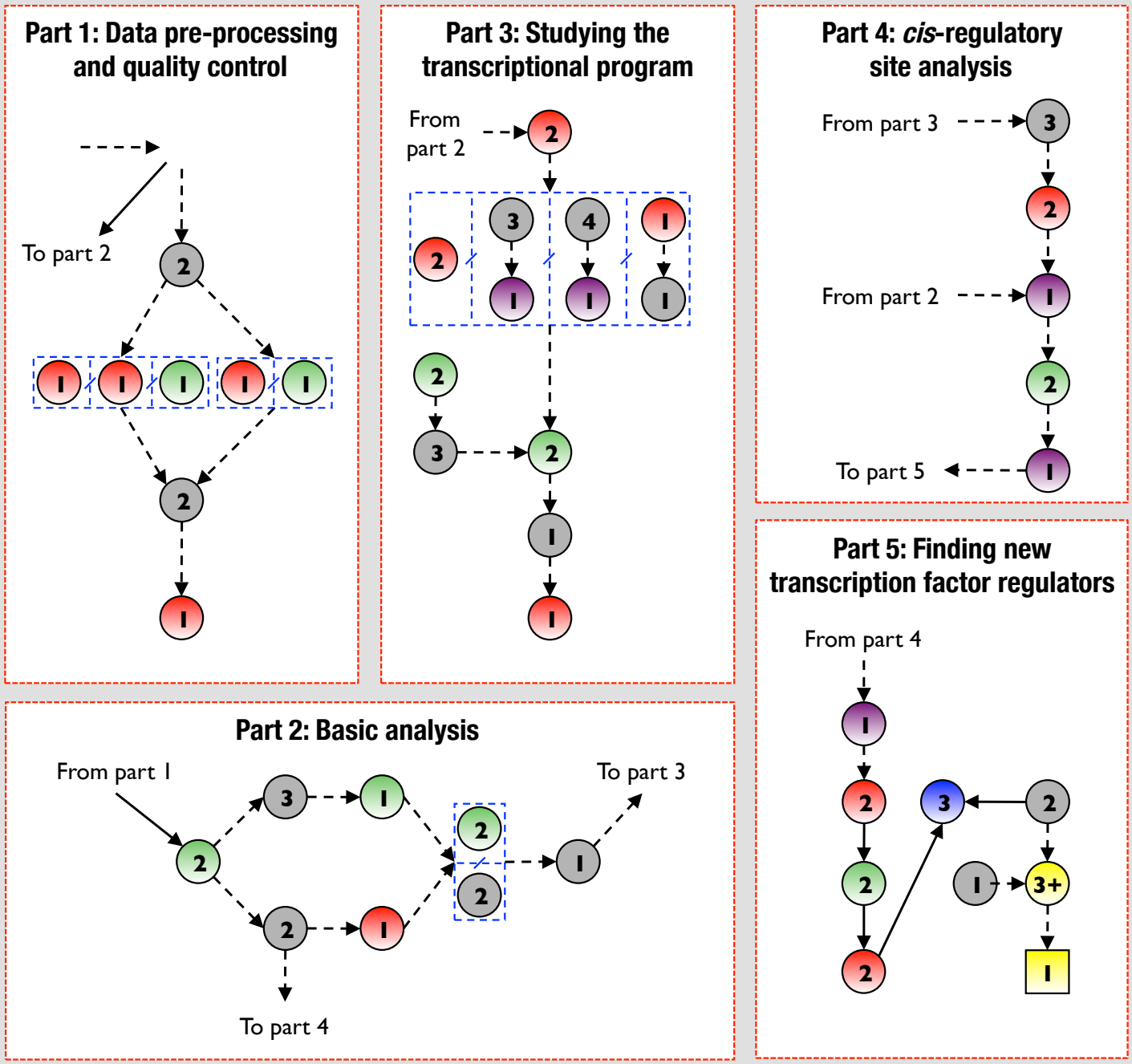
Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis

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

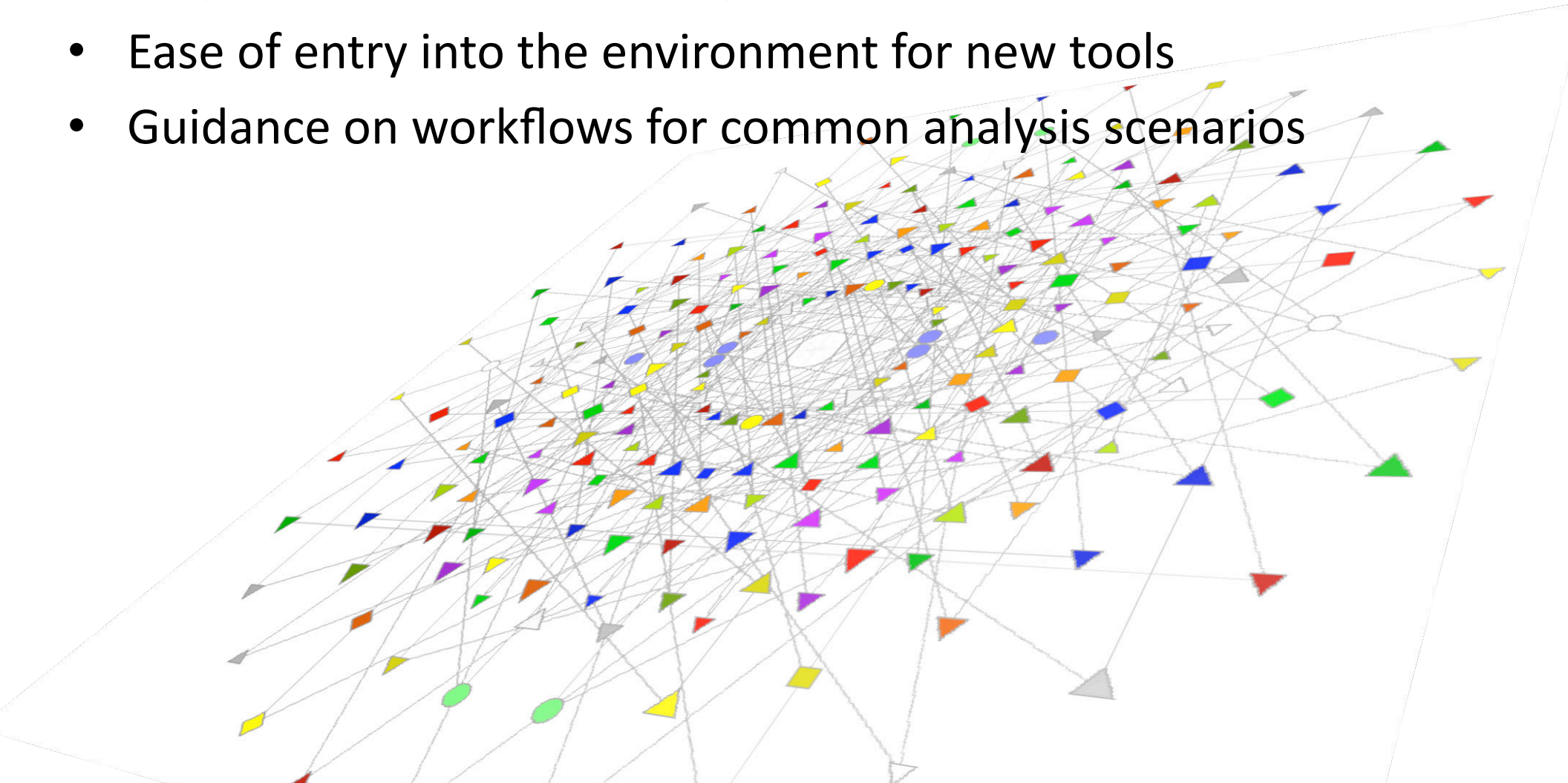


To do integrative research, you need to move data between many tools



We need a connection layer between tools

- Support for all types of resource: Web-based, desktop, etc.
- Automatic conversion of data formats between tools
- Easy access to data from any location
- Ease of entry into the environment for new tools
- Guidance on workflows for common analysis scenarios



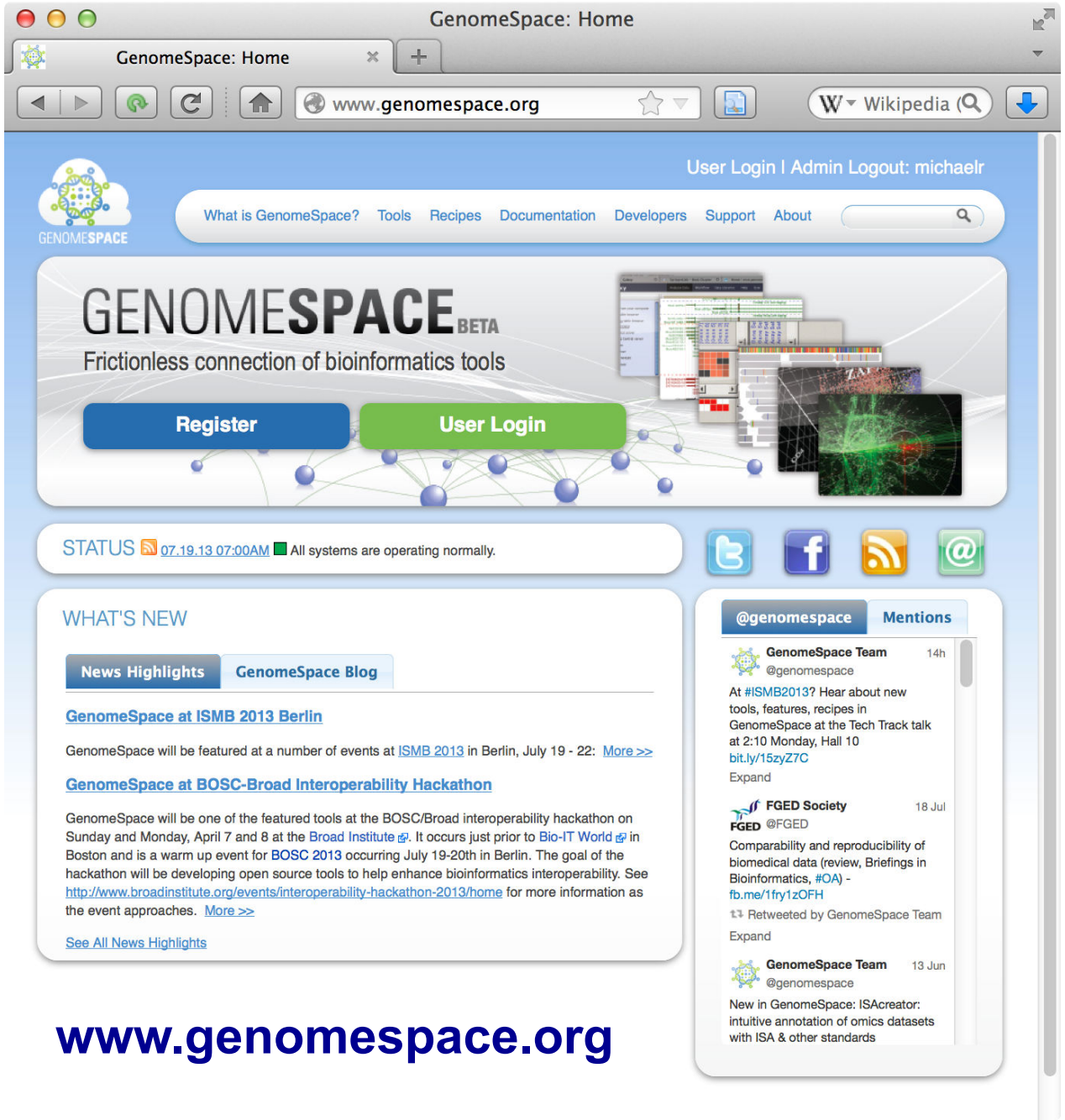
Twitter Founder Reveals Secret Formula for Getting Rich Online

BY RYAN TATE 09.30.13 9:30 AM

"The key to making a fortune online …
is to remove extra steps from common activities"
– Ev Williams

GENOMESPACE

An online community to share diverse computational tools



www.genomespace.org

Driving Biological Projects

lincRNAs
Cancer stem cells
Patient Stratification

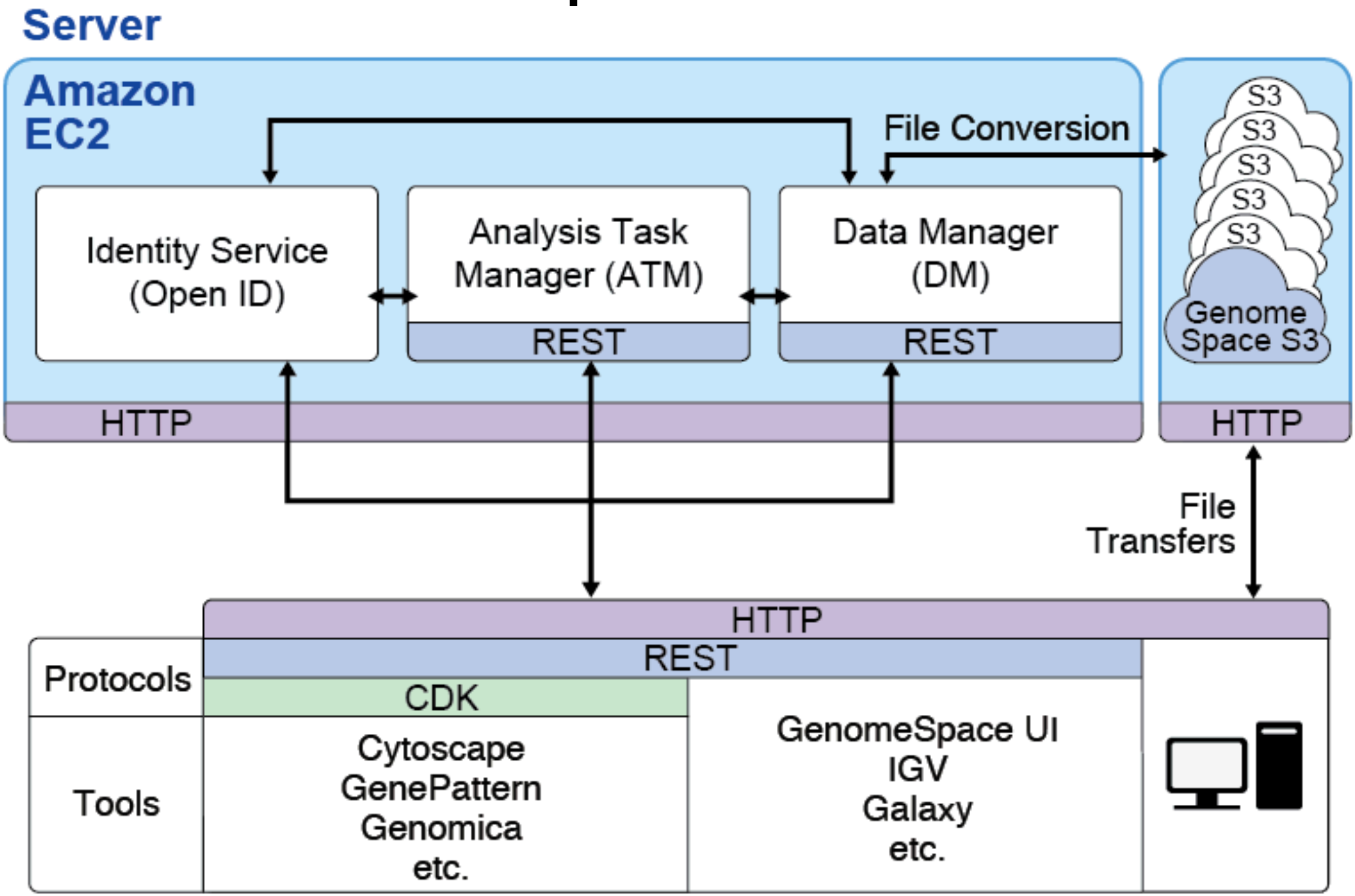
Seed Tools

Cytoscape
Galaxy
GenePattern
Genomica
IGV
UCSC Browser

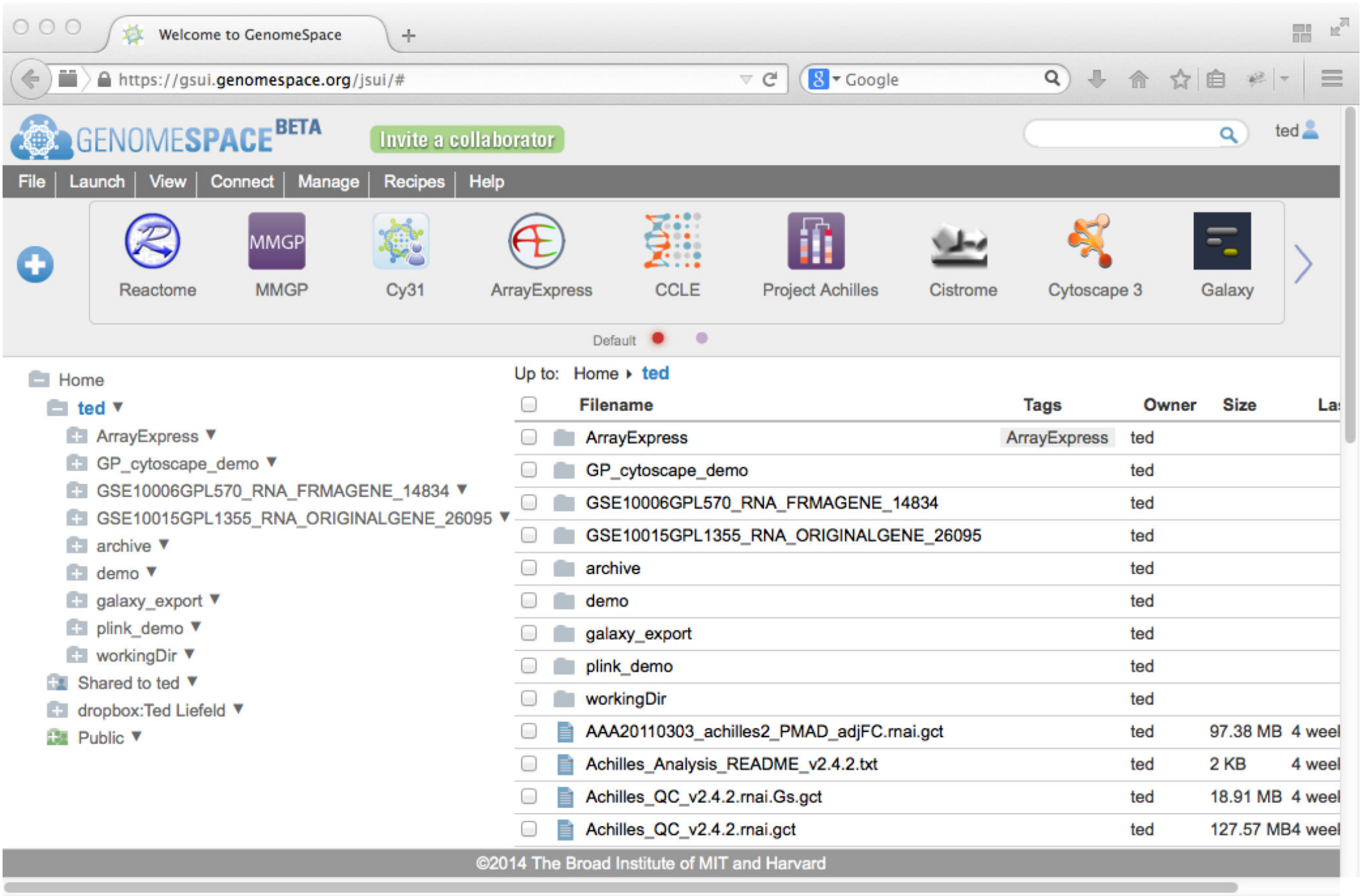
Analysis Recipes

Integrative analysis
Data transfer
etc.

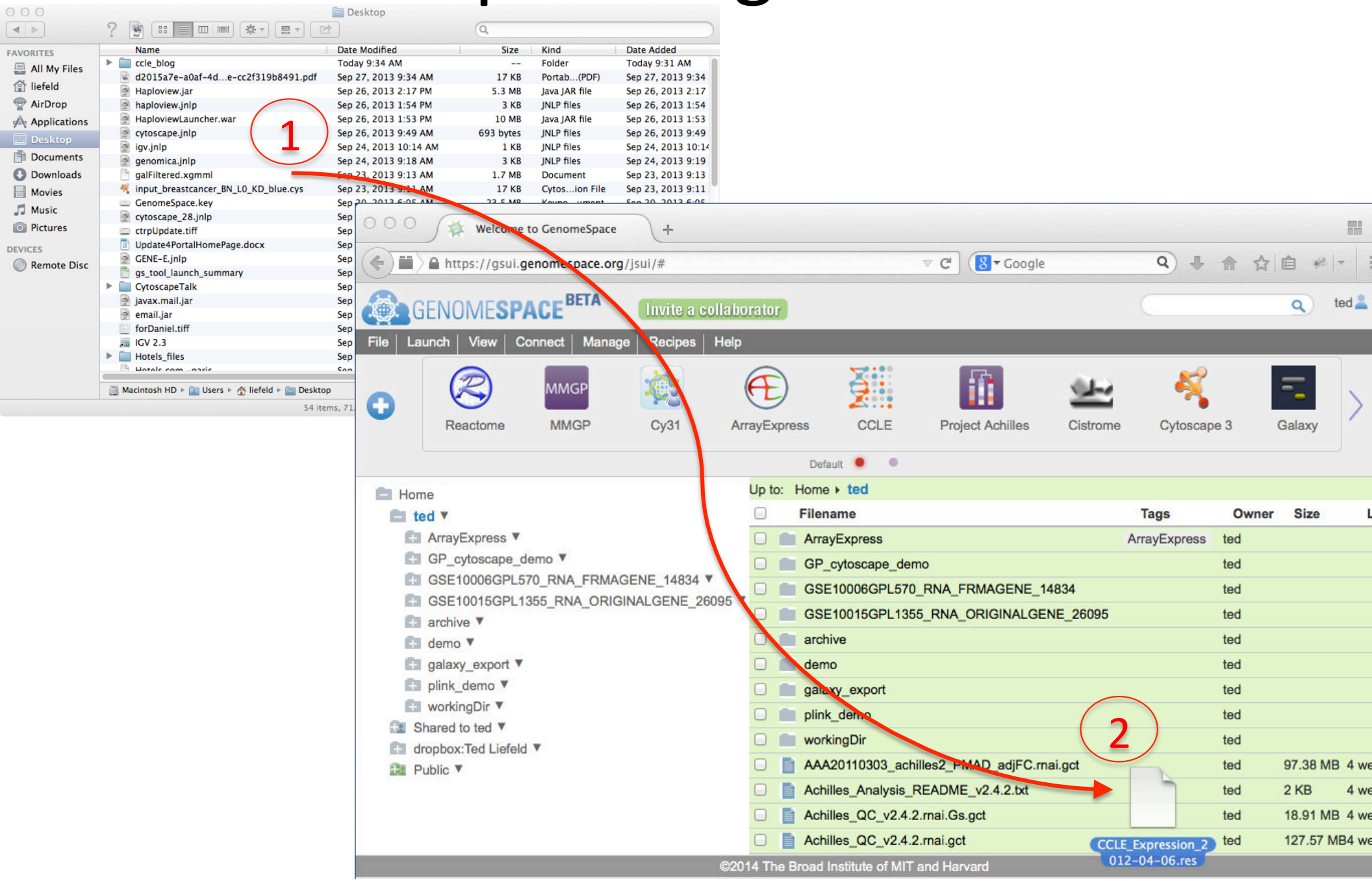
GenomeSpace Architecture



GenomeSpace UI



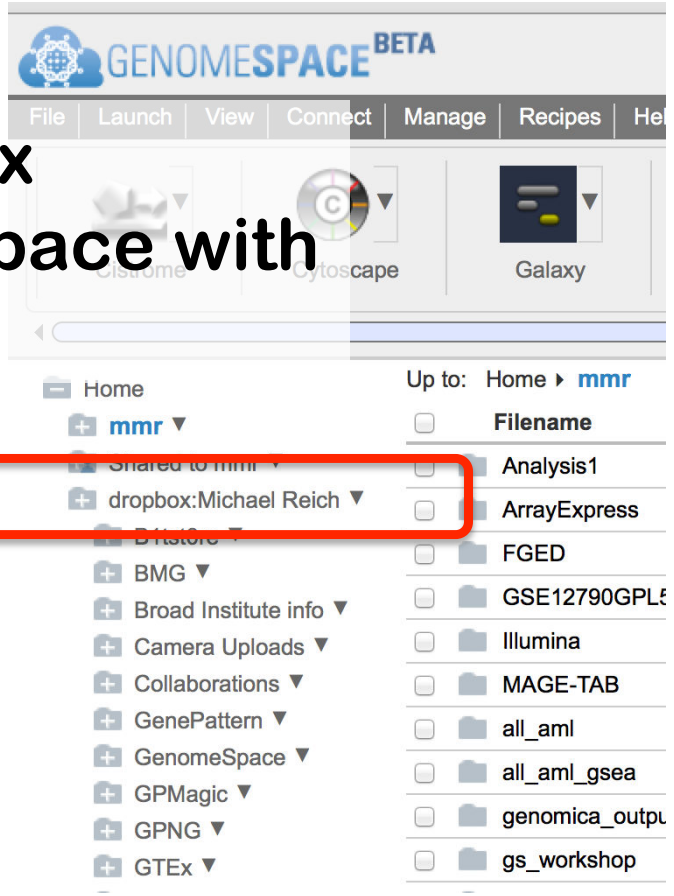
Uploading Data



Connection to cloud-based storage systems

Use GenomeSpace S3 storage or add your own Amazon account

Connect your Dropbox account to GenomeSpace with one click



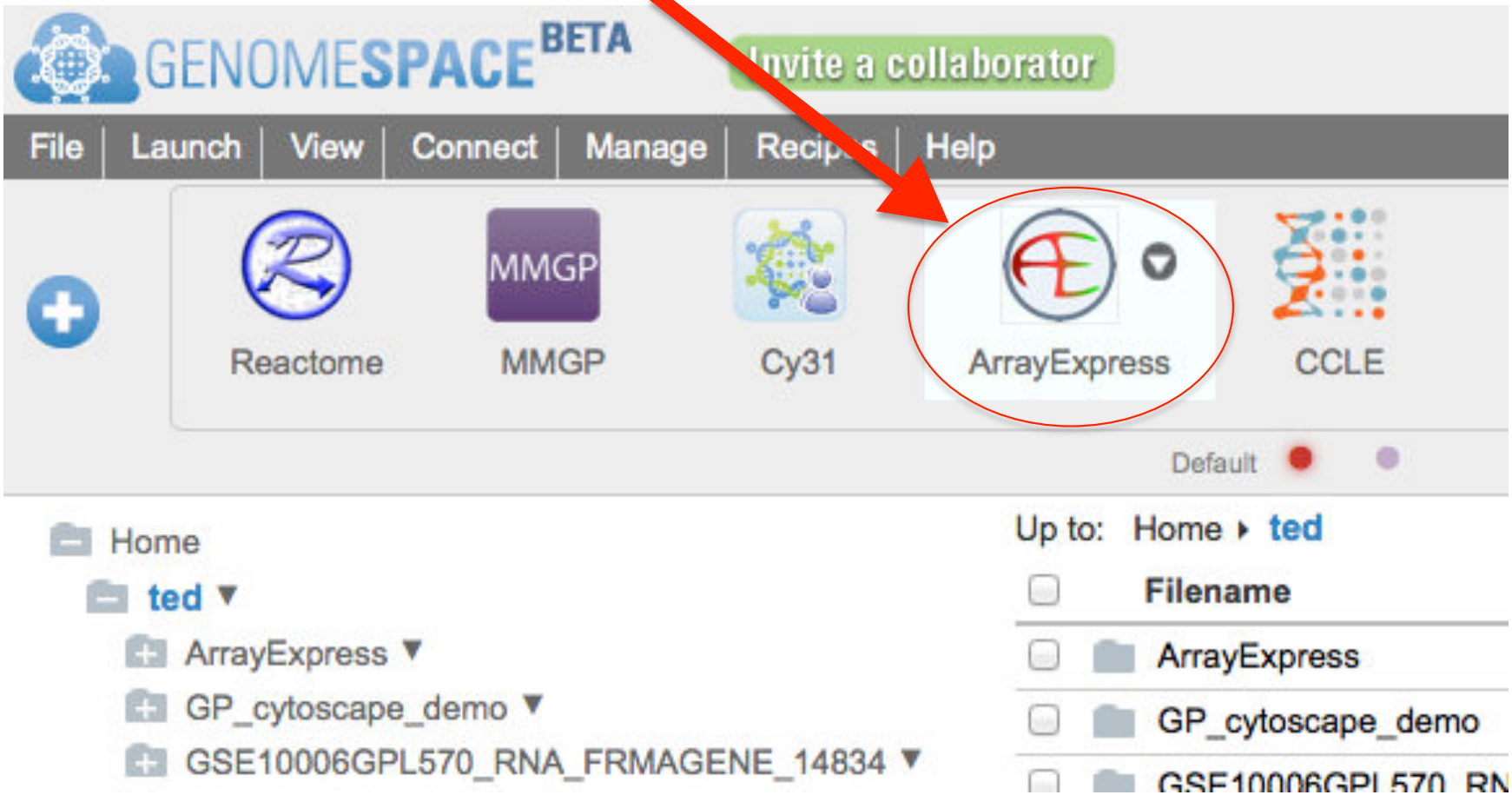
in development

openstack CLOUD SOFTWARE

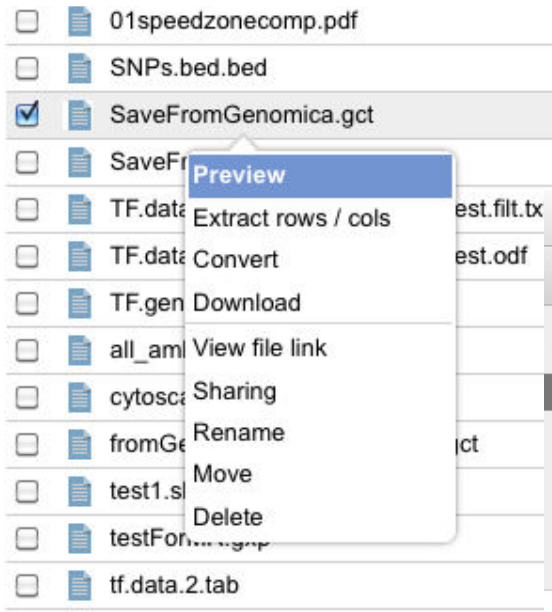
Google Drive

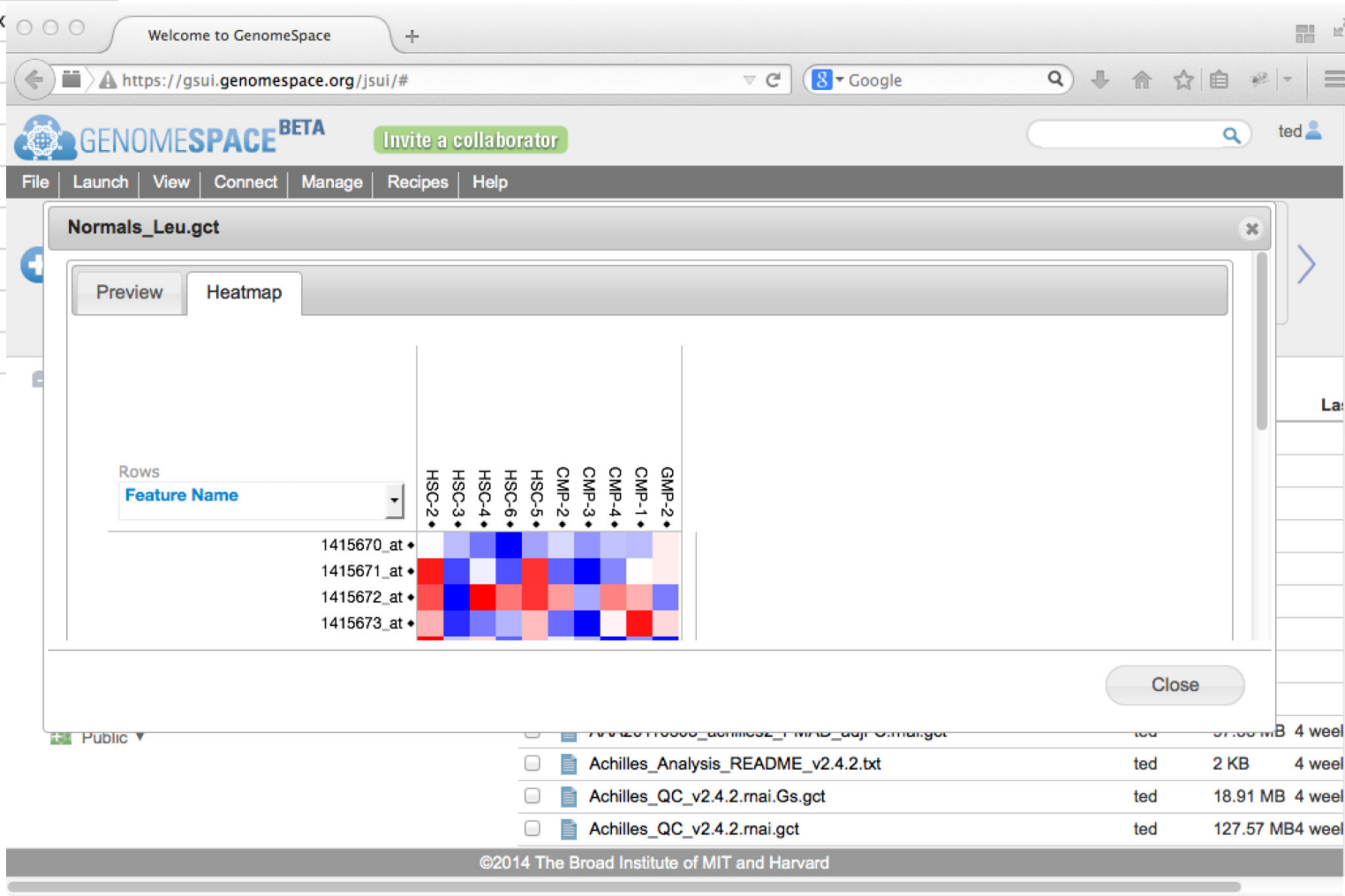
Launching tools

Click on the tool's icon

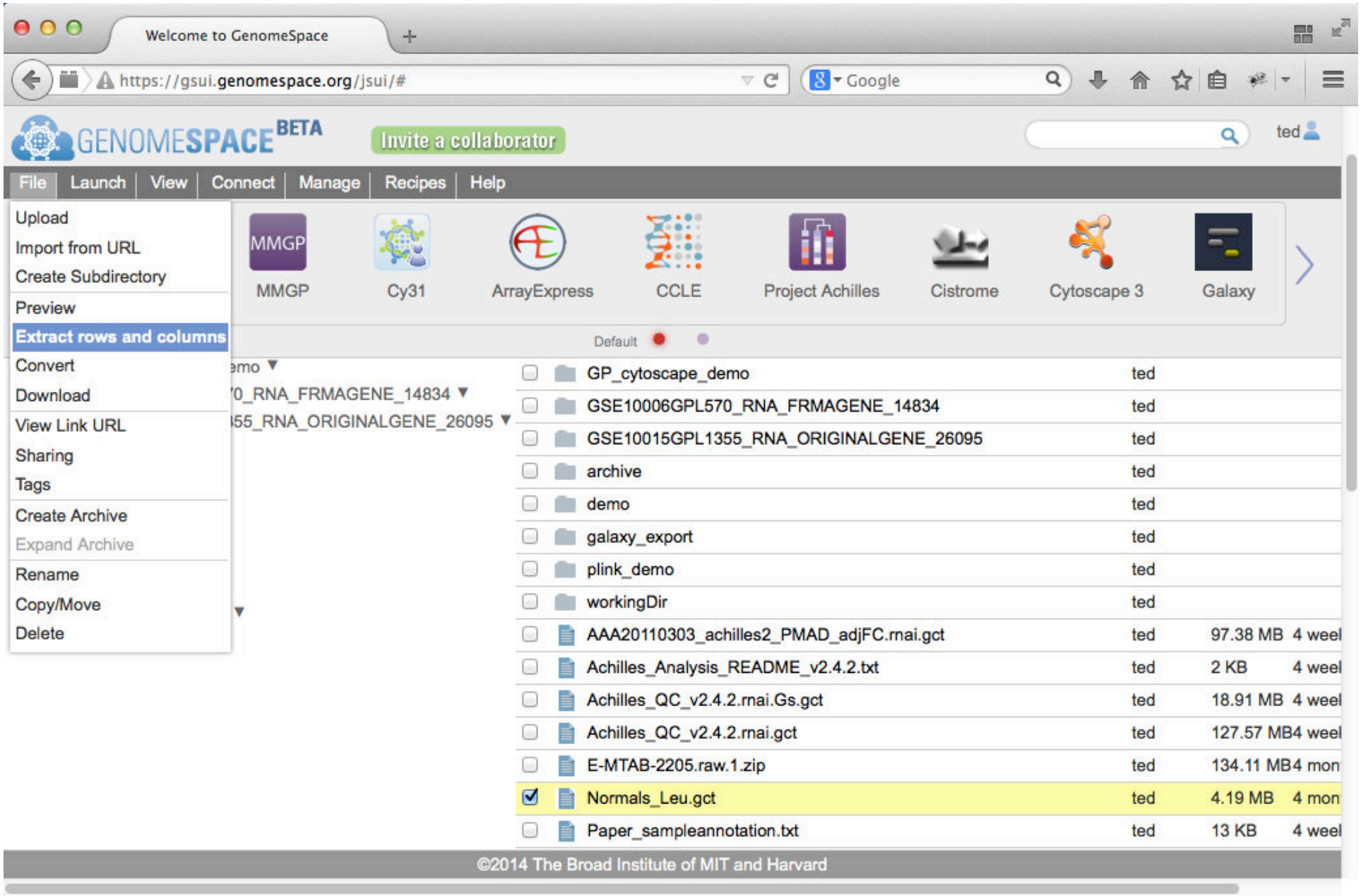


File Preview



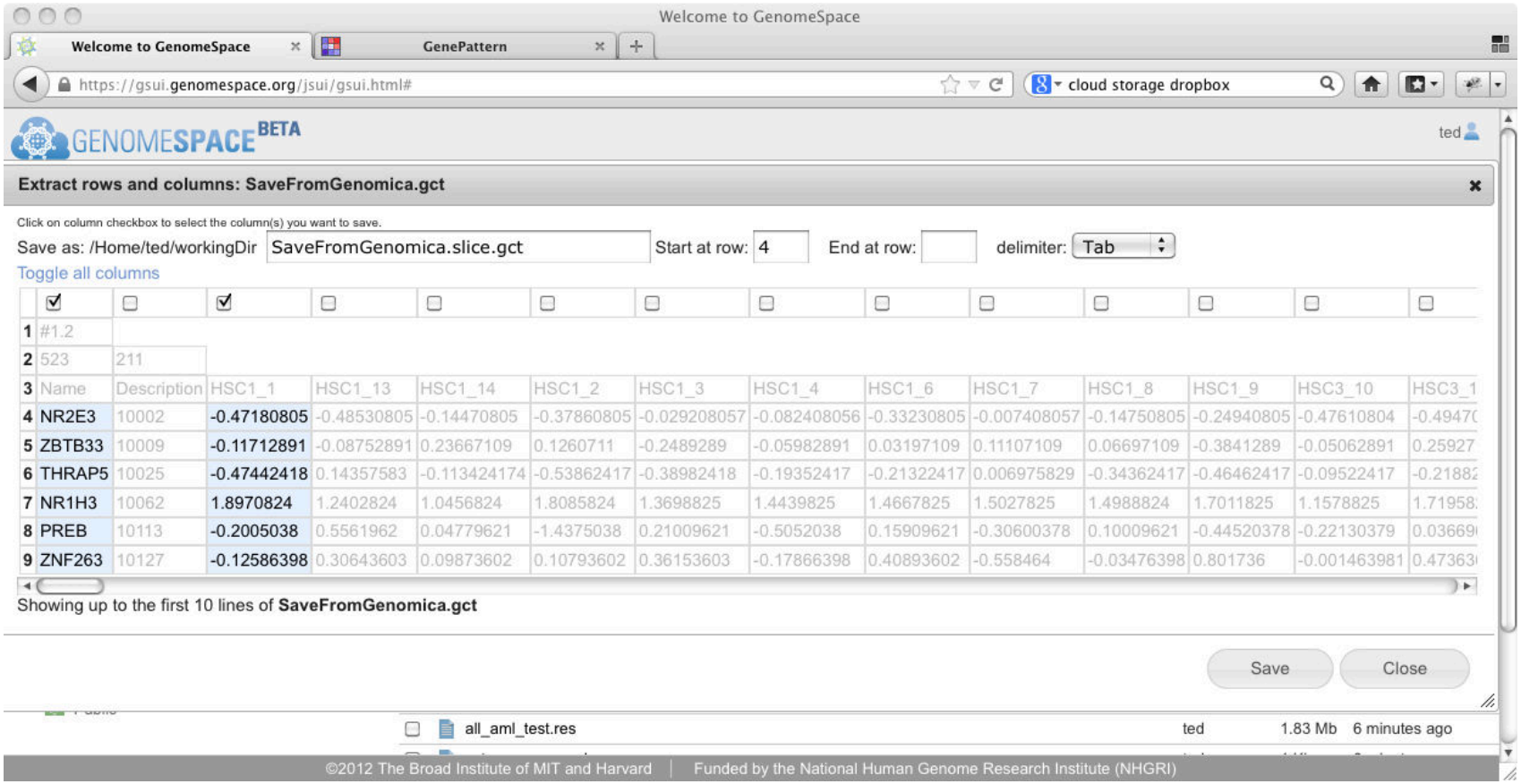


Extracting Rows and/or Columns



Extracting rows and/or columns

- Check the columns you want to include
- Provide a first (and optionally last) row index to include
- Edit the file name and ‘Save’

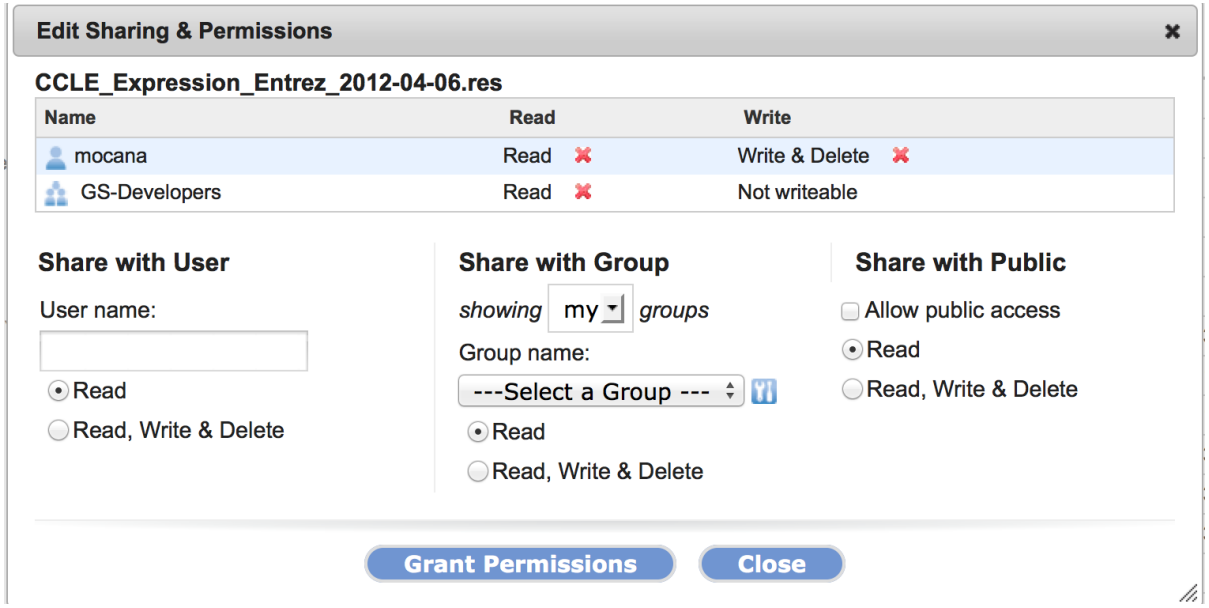


Sharing with others

- Sharing files with
  - Individuals

  - Groups

- Sharing links
  - With other GenomeSpace users
  - To people without GenomeSpace accounts



https://dm.genomespace.org/datamanager/file/Home/ted/CCLE_Expression_Entrez_2012-04-06.res

GenomeSpace-enabled tools, 7/2012

Galaxy

Cytoscape 2.8

GenePattern

IGV

UCSC Table Browser

Genomica

InSilicoDB

Cistrome

GenomeSpace-enabled tools, 7/2014

- Galaxy
- Cytoscape 2.8
- GenePattern
- ArrayExpress
- ISACreator
- Synapse
- GiTools
- Cytoscape 3
- MMGP
- CCLE
- IGV
- UCSC Table Browser
- Genomica
- InSilicoDB
- Cistrome
- GeWorkbench
- MSigDB
- Reactome
- Project Achilles

GenomeSpace-enabled tools, 7/2014

In development

cBio Portal *(Memorial Sloan Kettering Cancer Center)*

PLINK & HaploView *(Broad Institute/MGH)*

Synapse

GiTools

Cytoscape 3

MMGP

CCLE

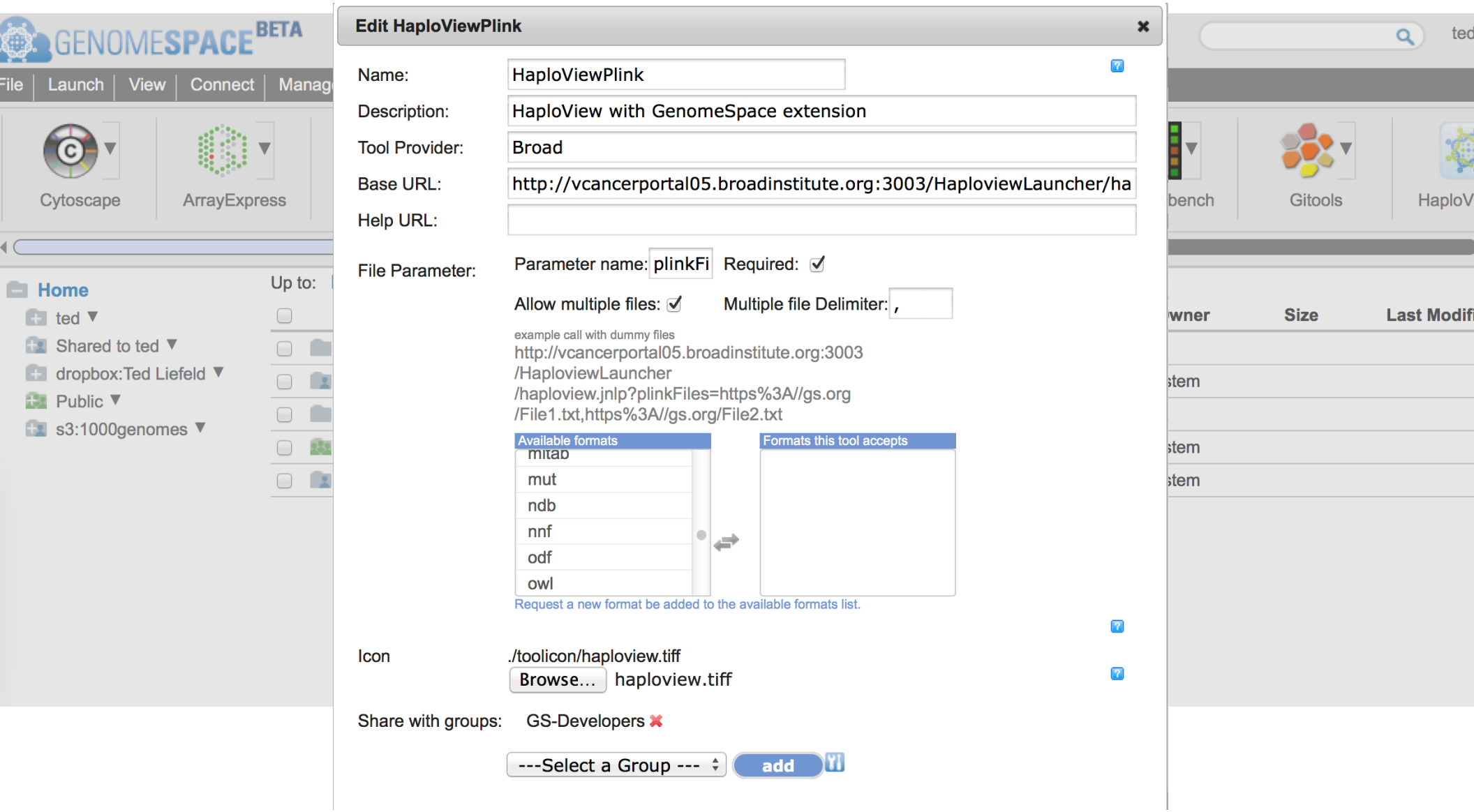
GeWorkbench

MSigDB

Reactome

Project Achilles

Or add your own tools

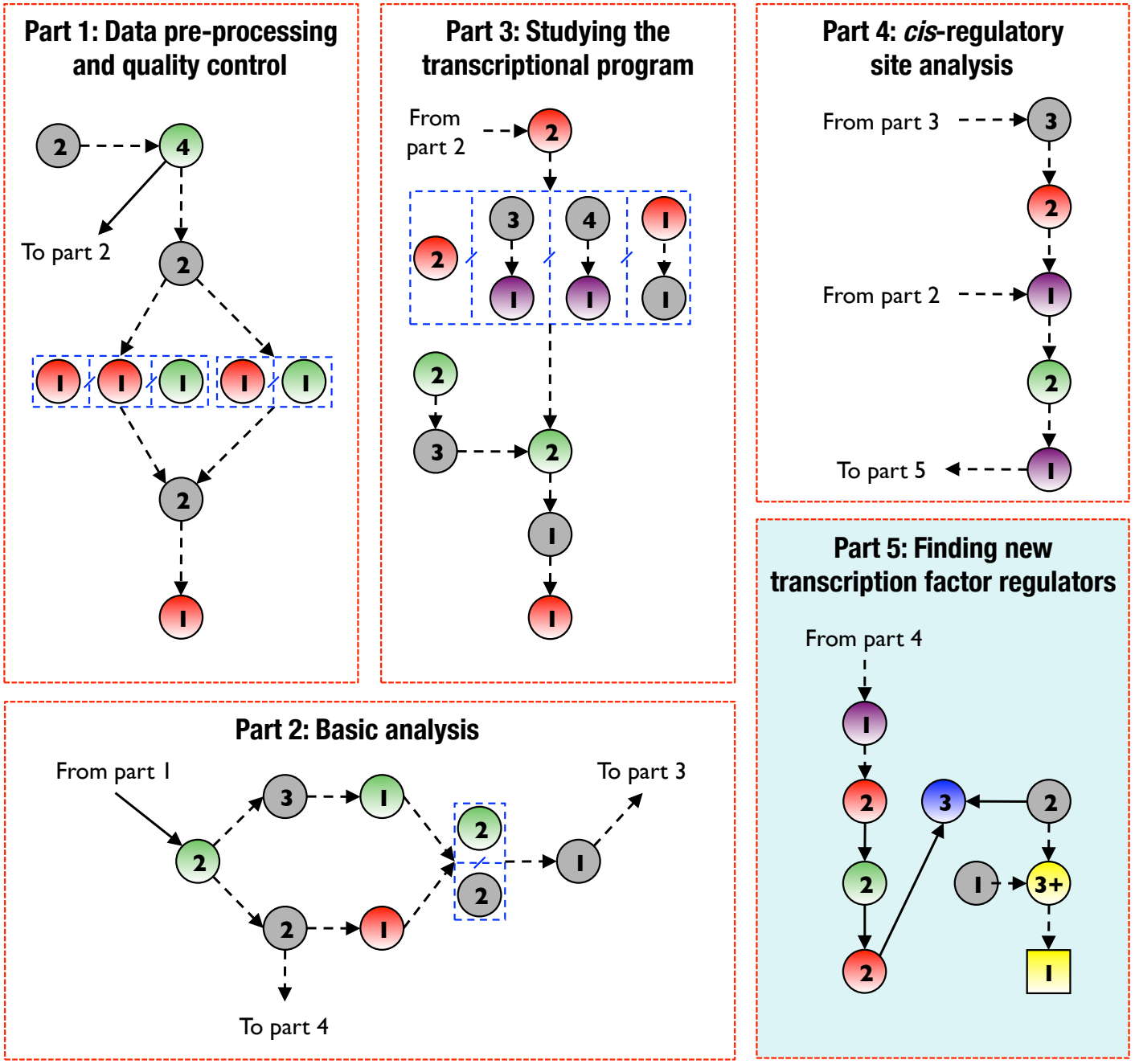


Automatic data transformation

From	To	Description
gct	Cytoscape ATTR	Converts GenePattern gct file to Cytoscape attribute format
Cytoscape ATTR	Cytoscape GeneMania attr	Converts attr to GeneMania attr
res	genomicatab	Converts res to Genomica tab format
adj	xgmml	Converts adjacency file to XGMML
gmt	genomicatab	Converts from gmt to Genomica tab using Perl script
gct	gxp	Converts gct stream to gxp format
res	gxp	Converts res stream to gxp format
genomicatab	gct	Converts Genomica Tab format to gct format
odf	Cytoscape ATTR	Converts GenePattern CMS results ODF file to Cytoscape attribute format
gxp	gct	Converts gxp stream to gct format
gct	genomicatab	Converts gct to Genomica tab format
gct	geneset.tab	Converts gct to Genomica geneset.tab
gct	geWorkbench exp	Converts gct file to exp file.
res	geneset.tab	Converts res to Genomica geneset.tab

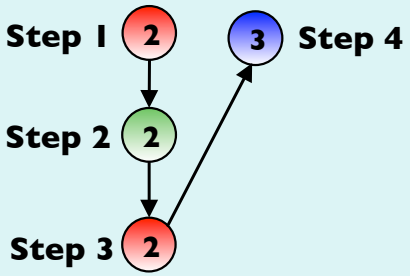
Data transformations added upon request and by user contribution

Studying the regulatory control of human hematopoiesis Using GenomeSpace



Part 5: Finding new transcription factor regulators

From part 4



Find the datasets in GenomeSpace

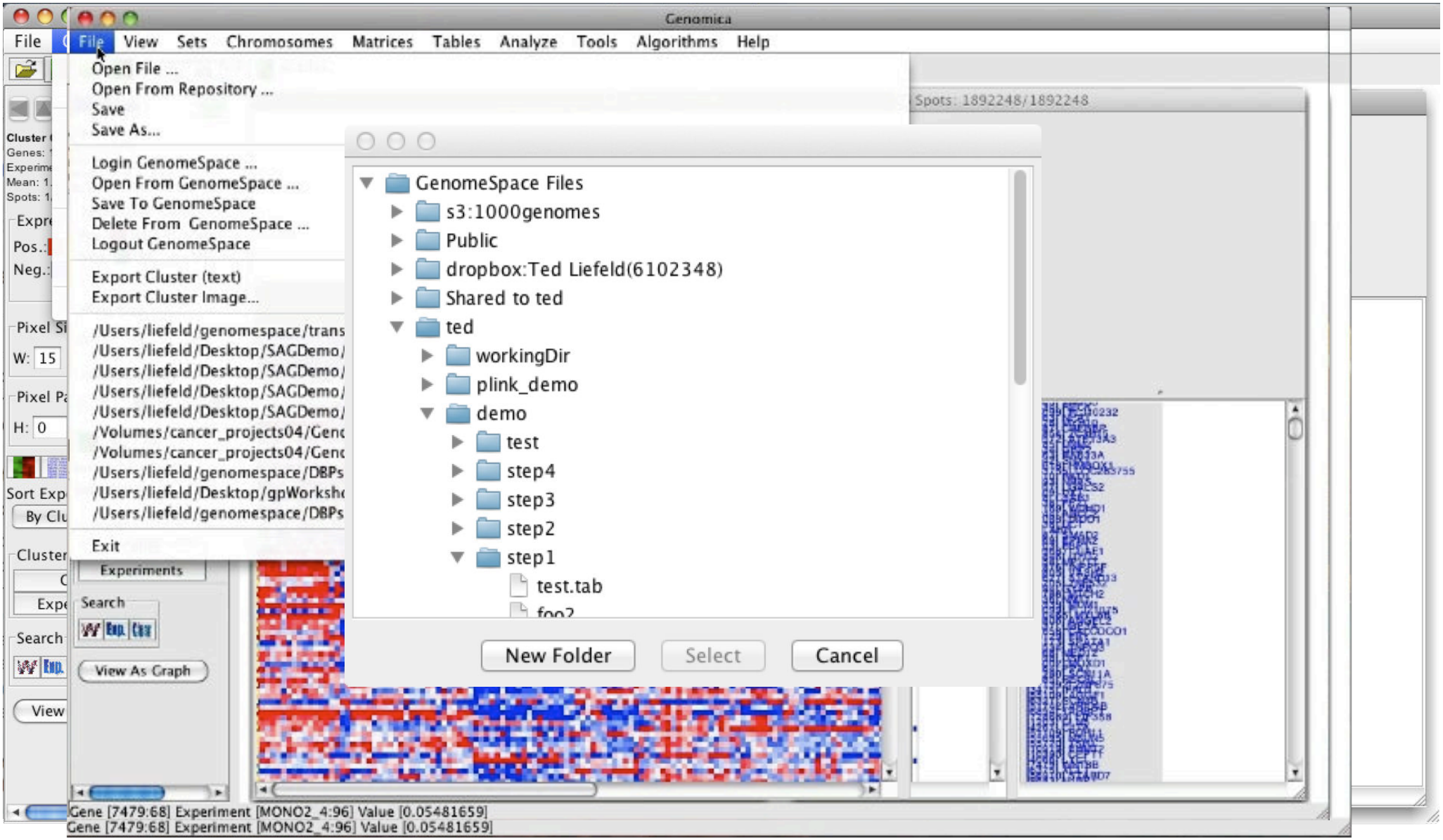
1 In Genomica, load expression data and transcription factor gene set, then extract transcription factor expression dataset

2 In GenePattern, run ComparativeMarkerSelection to find differentially expression transcription factors

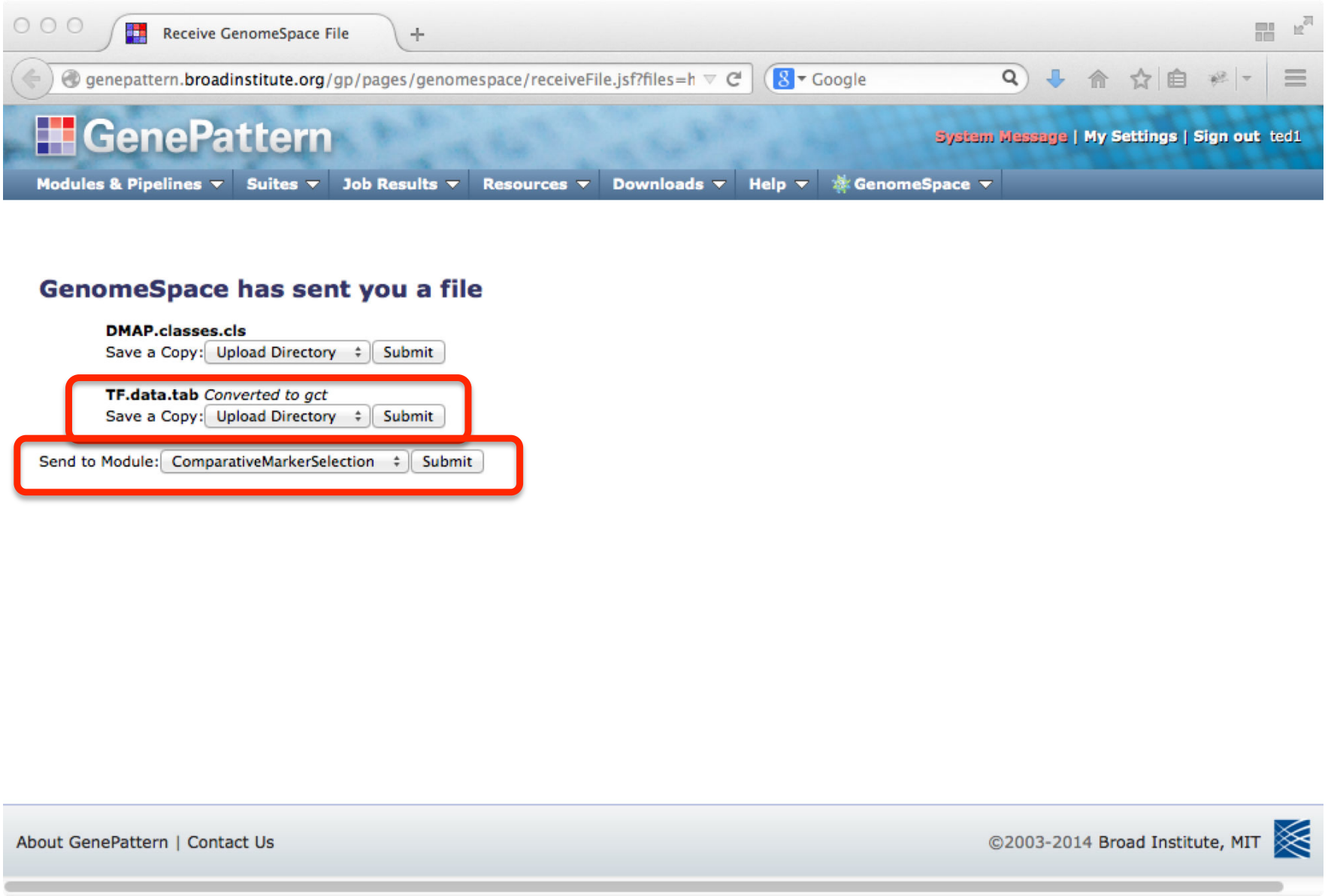
3 In Genomica, run Module Networks to get a list of regulators

4 In IGV, view loaded BED annotation tracks of:
1) transcription factors
2) SNPs
3) linkage regions

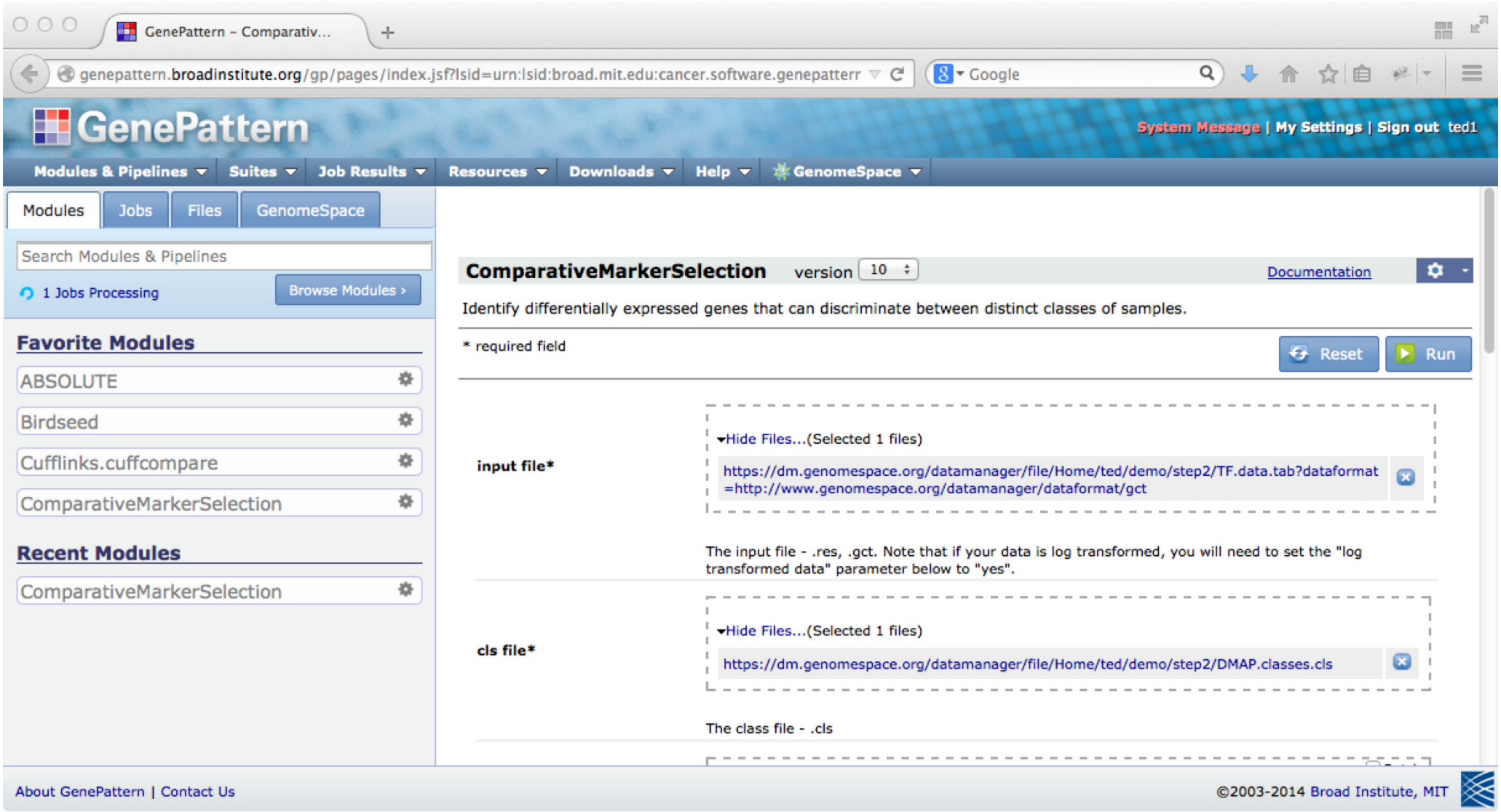
Step 1: Create transcription factor dataset in Genomica and save to GenomeSpace



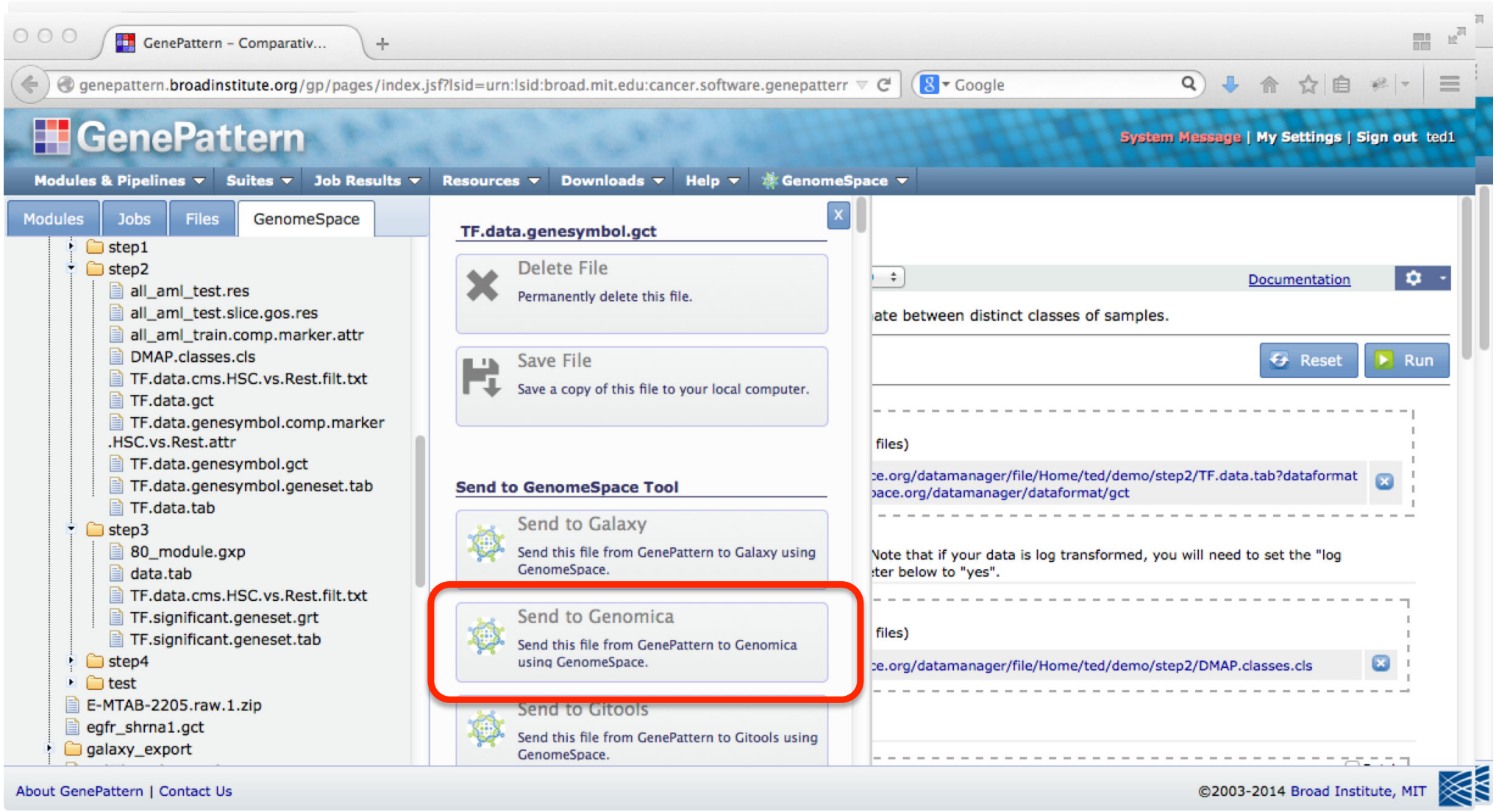
Step 2: Send transcription factor datasets into GenePattern



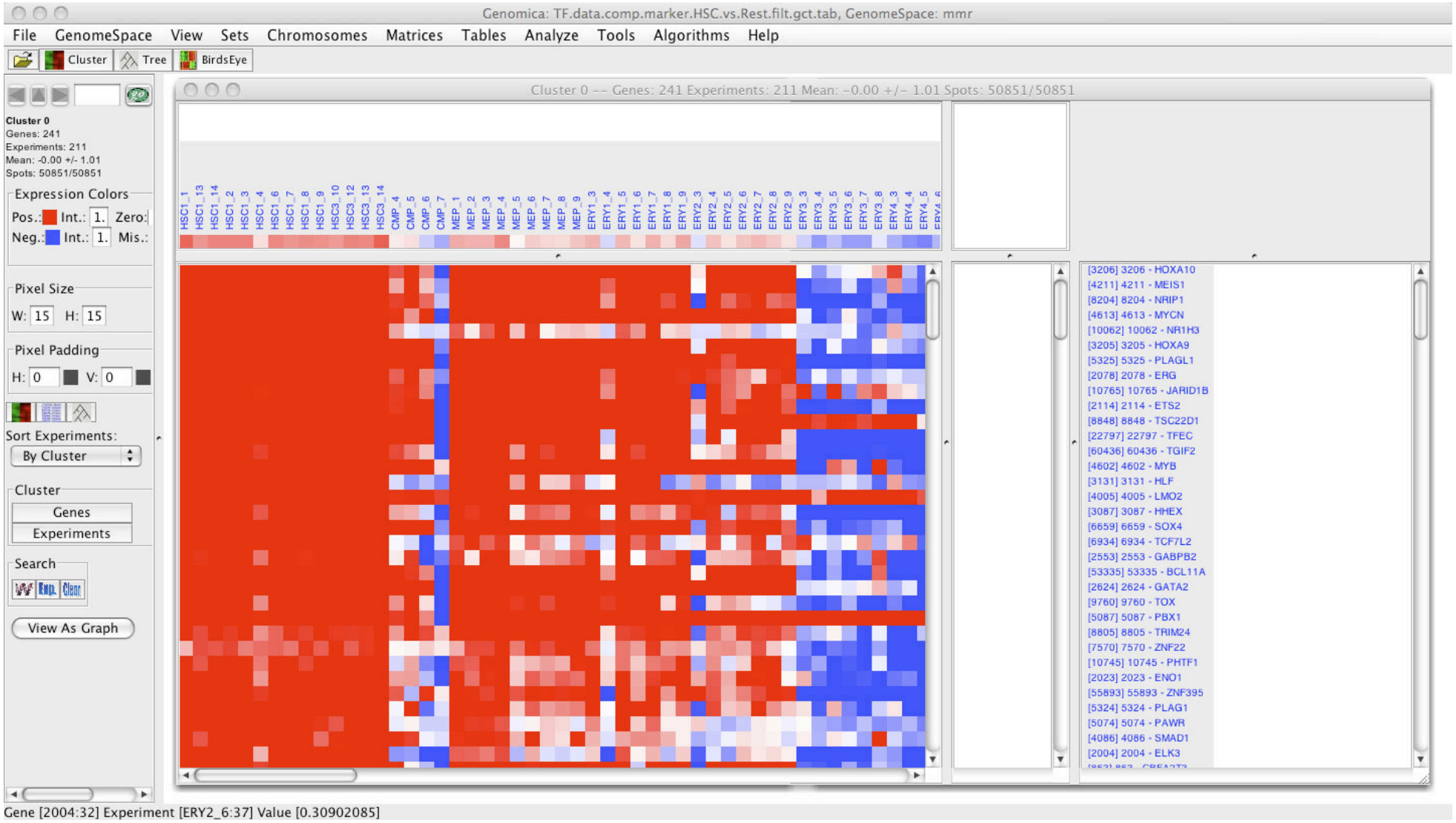
Step 2: Perform differential expression analysis in GenePattern



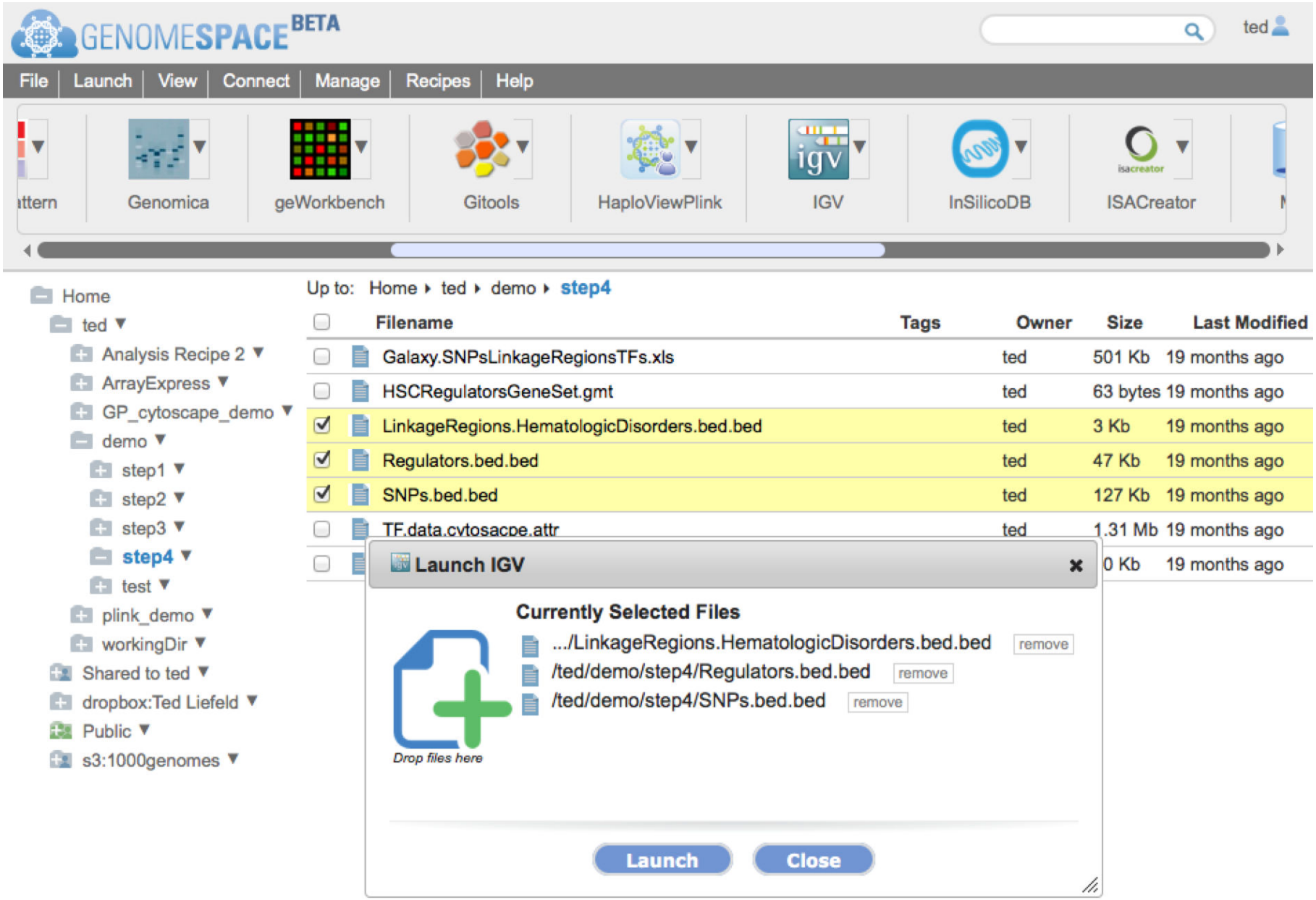
Step 3: Send differentially expressed genes to Genomica



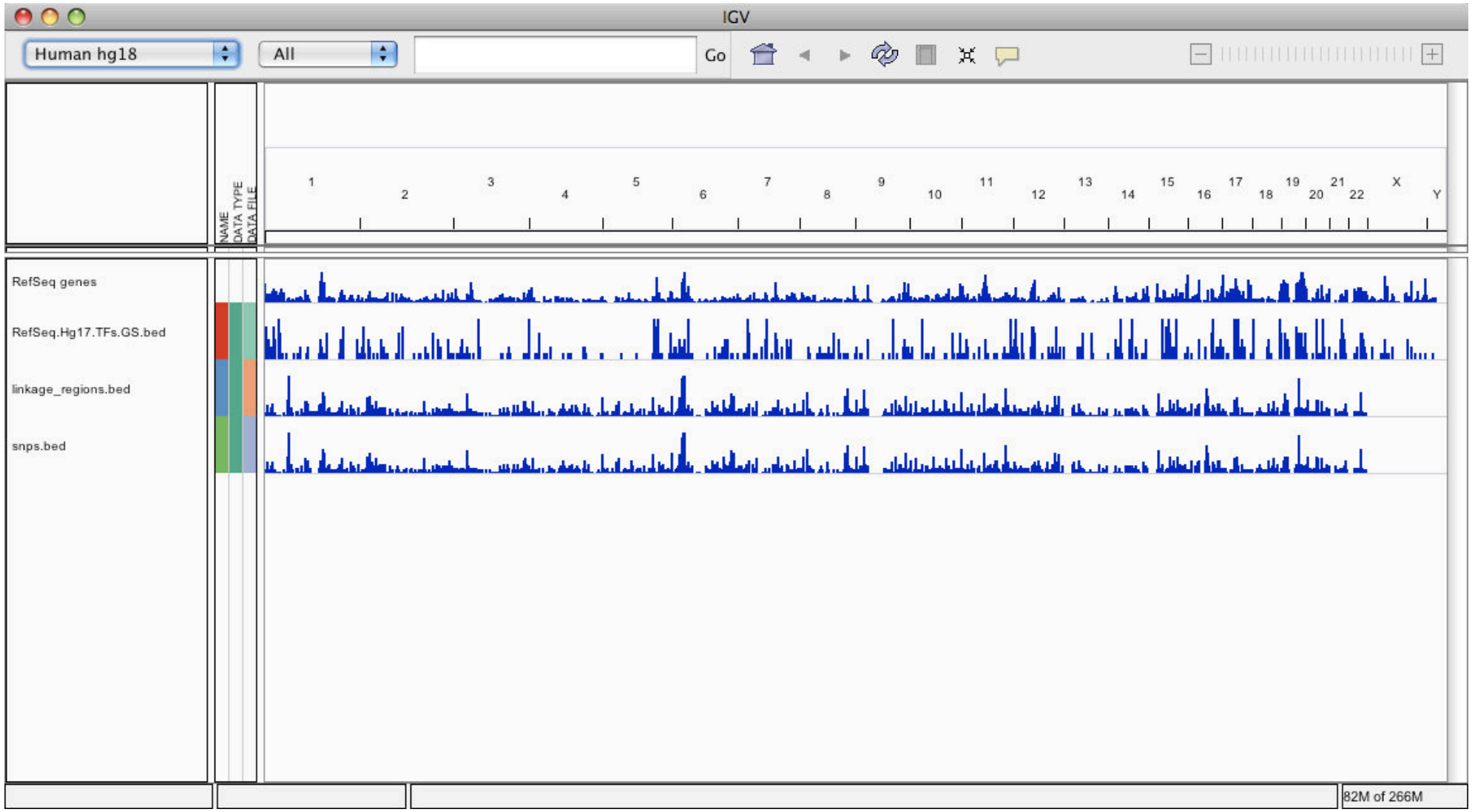
Step 3: perform module network analysis in Genomica



Step 4: Visualize regulators with known SNPs and linkage regions



Step 4: Visualize regulators with known SNPs and linkage regions



How about a non-retrospective example?

Cancer Discov. 2013 Mar;3(3):308-23. doi: 10.1158/2159-8290.CD-12-0418. Epub 2013 Feb 21.

Targeting MYCN in neuroblastoma by BET bromodomain inhibition.

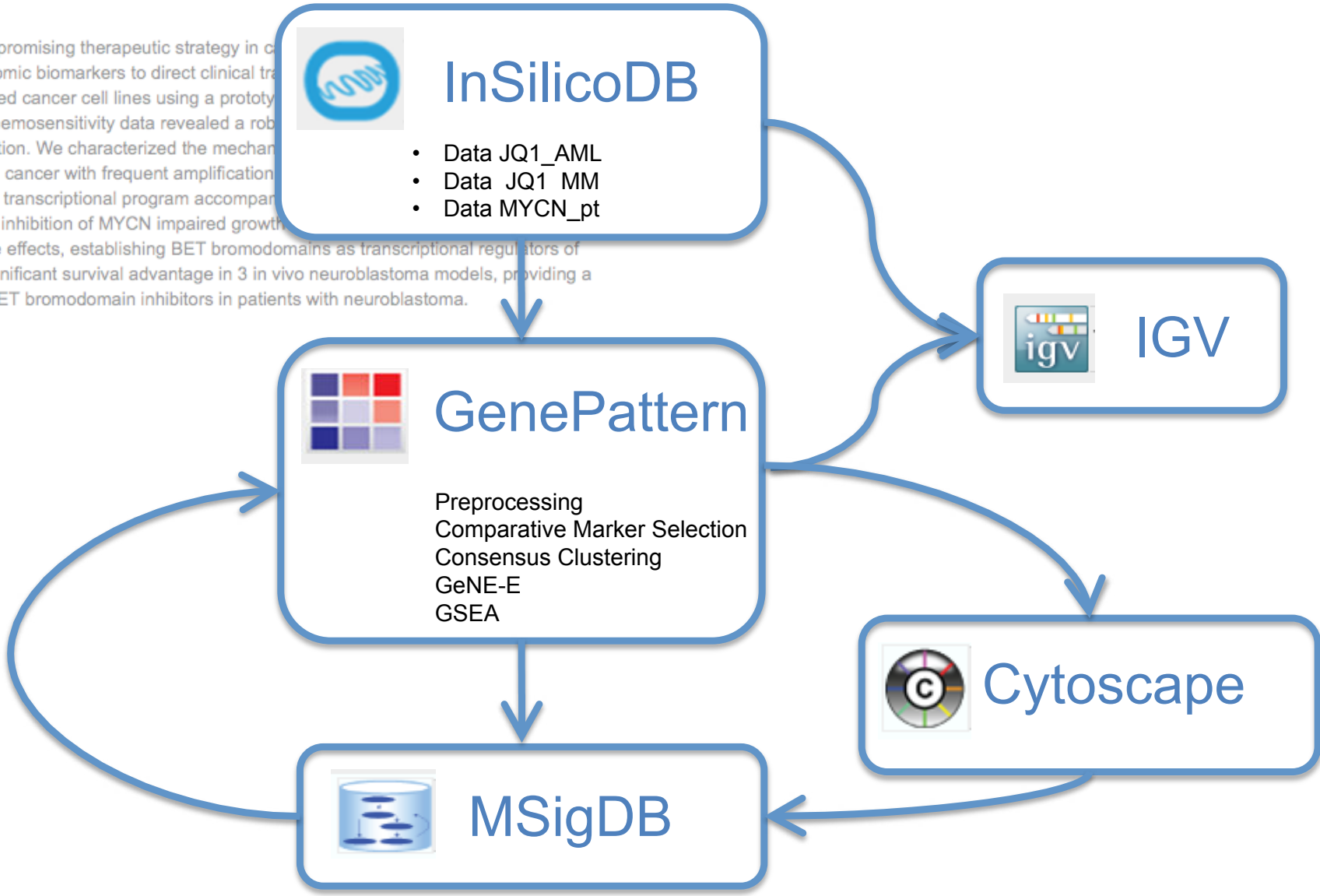
Puissant A, Frumm SM, Alexe G, Bassil CF, Qi J, Chanthery YH, Nekritz EA, Zeid R, Gustafson WC, Greninger P, Garnett MJ, McDermott U, Benes CH, Kung AL, Weiss WA, Bradner JE, Stegmaier K.

Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.

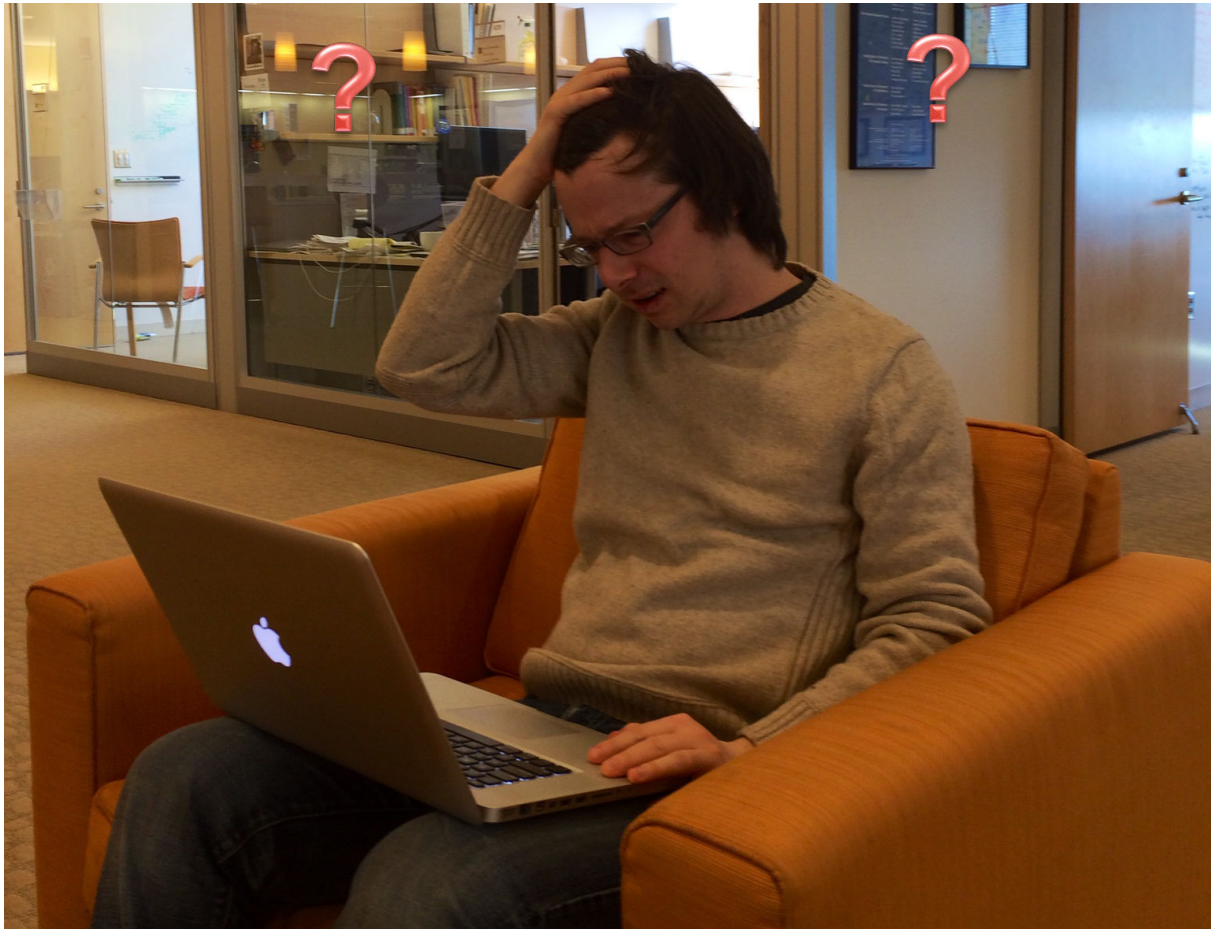
Abstract

Bromodomain inhibition comprises a promising therapeutic strategy in c
malignancies. To date, however, genomic biomarkers to direct clinical tr
cell-based screen of genetically defined cancer cell lines using a prototy
Integration of genetic features with chemosensitivity data revealed a rob
and sensitivity to bromodomain inhibition. We characterized the mechan
finding in neuroblastoma, a childhood cancer with frequent amplification
showed downregulation of the MYCN transcriptional program accompan
Functionally, bromodomain-mediated inhibition of MYCN impaired growth
BRD4 knockdown phenocopied these effects, establishing BET bromodomains as transcriptional regulators of MYCN. BET inhibition conferred a significant survival advantage in 3 in vivo neuroblastoma models, providing a compelling rationale for developing BET bromodomain inhibitors in patients with neuroblastoma.

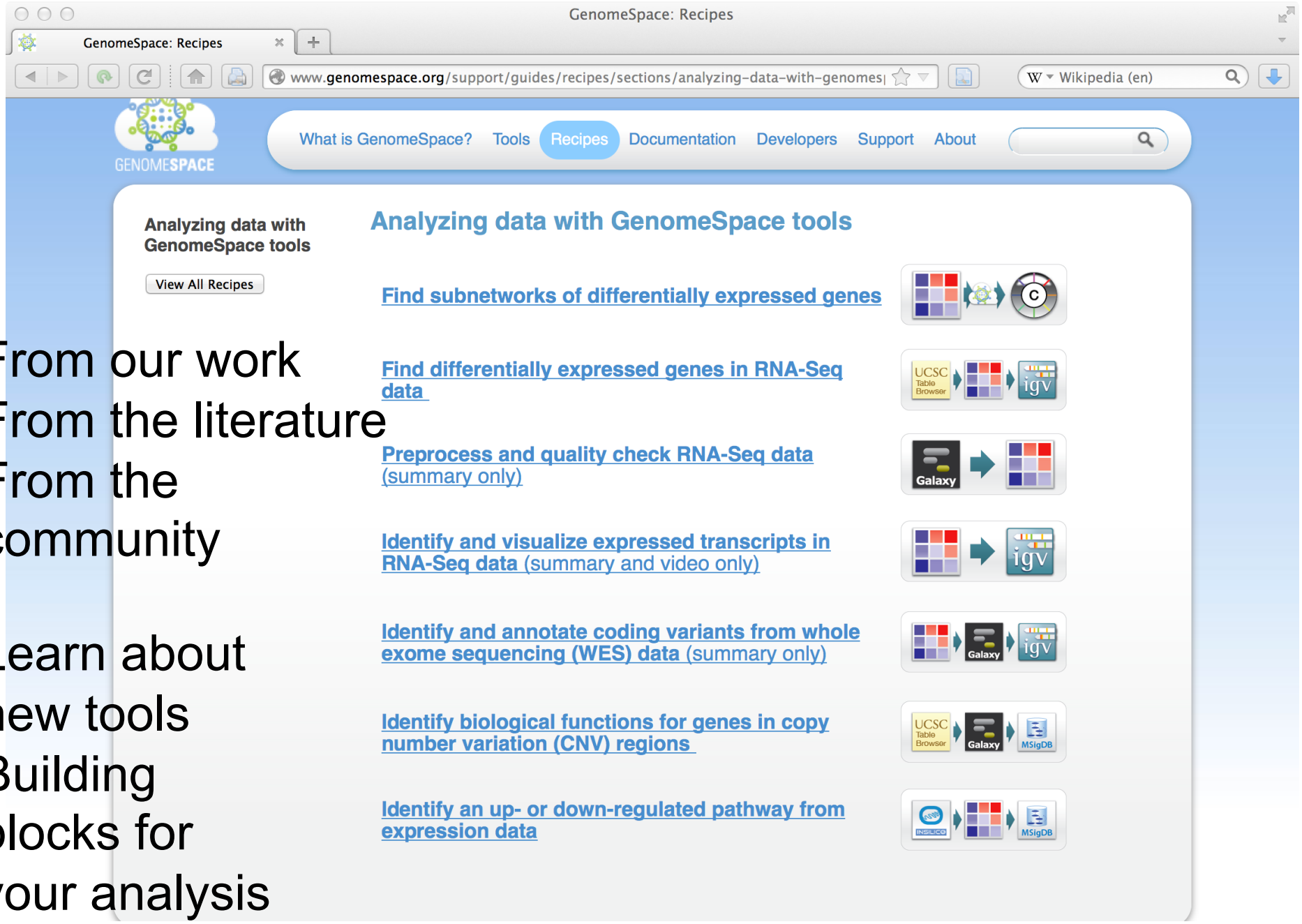
Comment in



How do I use it???



GenomeSpace Recipe Collection



1. From our work
2. From the literature
3. From the community

- Learn about new tools
- Building blocks for your analysis

Example Recipe: find subnetworks of differentially expressed genes.

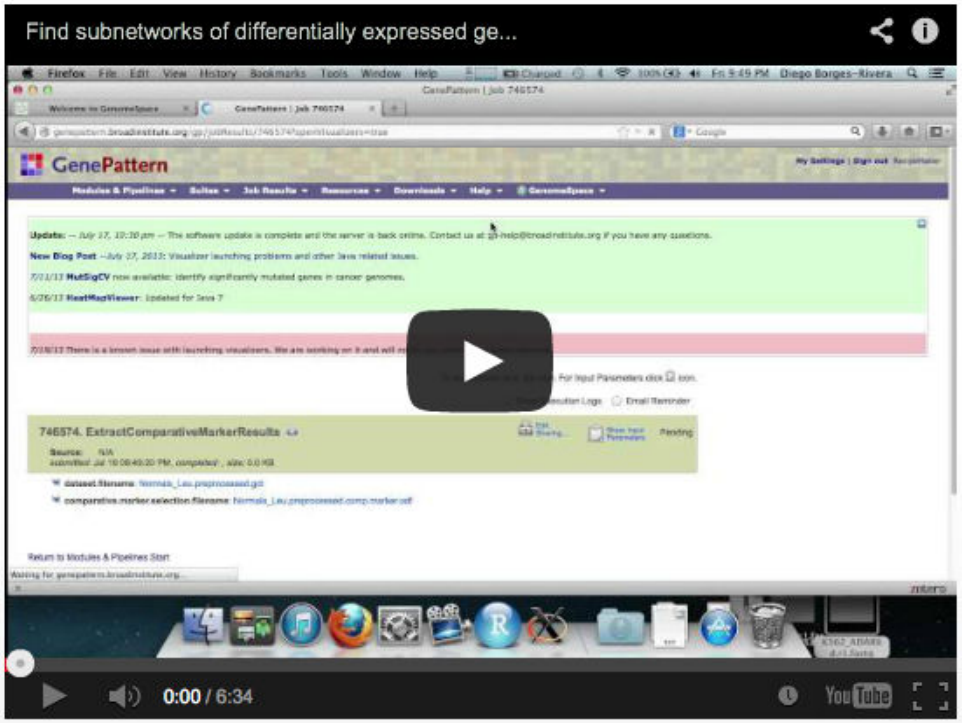
Find subnetworks of differentially expressed genes

1. Go to GenePattern from GenomeSpace
2. Filter out rows having less than 3x expression variation
3. Find differentially expressed genes
4. Select the top 50 genes by rank
5. Get a gene list
6. Get the network of interacting proteins
7. Navigate and explore the protein-protein network
8. Find subnetworks

View All Recipes

Workflow →

Find subnetworks of differentially expressed genes



← Video

Summary

Summary →

This recipe provides one method to finding subnetworks of interacting proteins that are differentially expressed between normal cells and tumor cells.

Input →

Input:

- A gene expression dataset, containing data from two different conditions or phenotypes, such as tumor and normal
- A class file identifying the classes of samples in the gene expression dataset

Steps and Tools →

Steps:

1. Find the list of the 50 top differentially expressed genes in a gene expression dataset using **PreprocessDataset**, **ComparativeMarkerSelection**, and **ExtractComparativeMarkerResults** (in **GenePattern**)
2. Find the protein-protein interaction subnetworks associated with these genes using **GeneMANIA** and **MCODE** (in **Cytoscape**)

Example Recipe: find subnetworks of differentially expressed genes.

Use case →

Detailed steps →

Recipe Details

We describe the detailed steps below in the context of an example, where we are using gene expression data for normal progenitor and leukemic samples from a study of transformation from committed progenitor cells to leukemia stem cells initiated by MLL-AF9 to find differentially expressed subnetworks. The example data is from mouse (*Mus musculus*) cell lines. Expression input data is in GenePattern's GCT format, and the two classes of cells are identified in GenePattern's CLS format. The example data files are available in the GenomeSpace Public folder (*Public>SharedData>Demos>Analysis Recipe 1*):

- `Normal_Leu.gct` (Gene expression data for normal *Mus musculus* progenitor and leukemic samples from a study of transformation from committed progenitor cells to leukemia stem cells initiated by MLL-AF9)
- `Normal_Leu.cls` (Class assignments -- tumor or normal -- for all the samples in the GCT file)

GenePattern: Find differentially expressed genes

1. Go to GenePattern from GenomeSpace

a. Log into GenomeSpace.
b. Click the GenePattern button in the GenomeSpace tool bar.
c. In the GenePattern window of the browser, click *Sign In Using GenomeSpace*.

Note: If you have not yet associated your GenomeSpace login with your GenePattern login, you will be asked to do so, or you can automatically generate a new GenePattern account that will be associated with your GenomeSpace login.

2. Filter out rows having less than 3x expression variation

The **PreprocessDataset** module trims out many of the rows of data for genes with expression variation less than the 3x (up- or down-regulated) threshold, and also normalizes the data.

a. In the *Modules & Pipelines* search box, start typing `PreprocessDataset`, and click the name when it appears in the options dropdown under the box.

Where we're at…

- Over 5000 registered users
- 19 public tools enabled + 15 private tools

  ~400 tool launches/week, ~1500 up/downloads, ~3TB user data
- 19 recipes posted

  –RNA-Seq, Microarray, CNV, Network data, etc.
- Publications

  –GenomeSpace manuscript in draft

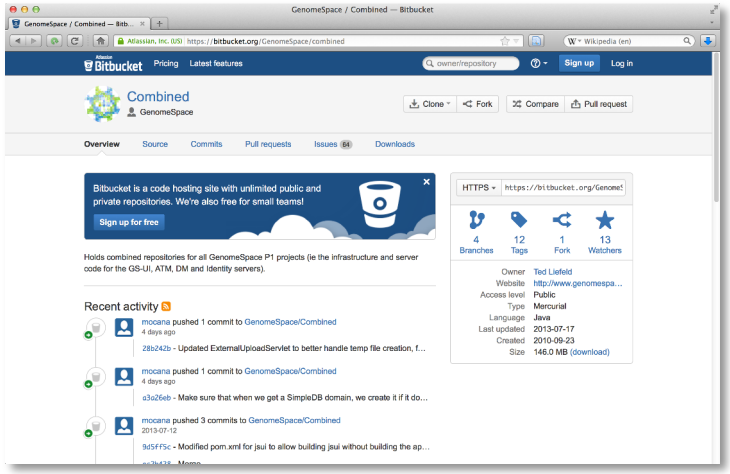
  –'Cytoscape: the network visualization tool for GenomeSpace workflows' July 2014, in F1000Research
- Exploring moving deployment from AWS to TACC and Bionimbus
- Developing new, dynamic, recipe system and new recipes

Explore GenomeSpace

sign up at www.genomespace.org

- **Researchers**
  - Analyze data
  - Request format converters
  - Contribute analysis recipes
- **Developers**
  - Add your tools
  - Contribute format converters
  - Build new infrastructure
  - Hook up your app to ours
- **Data providers**
  - Link your data

- **Cores**
  - Contribute and share recipes for your customers
  - Add your local tools



Fork us on Bitbucket

Acknowledgements

GenomeSpace Collaborators
Cytoscape: Trey Ideker Lab, UCSD
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

Broad Institute
Sara Garamszegi
Marco Ocana
Michael Reich
Jim Robinson
Thorin Tabor
Helga Thorvaldsdottir
Felix Wu

Jill Mesirov, PI

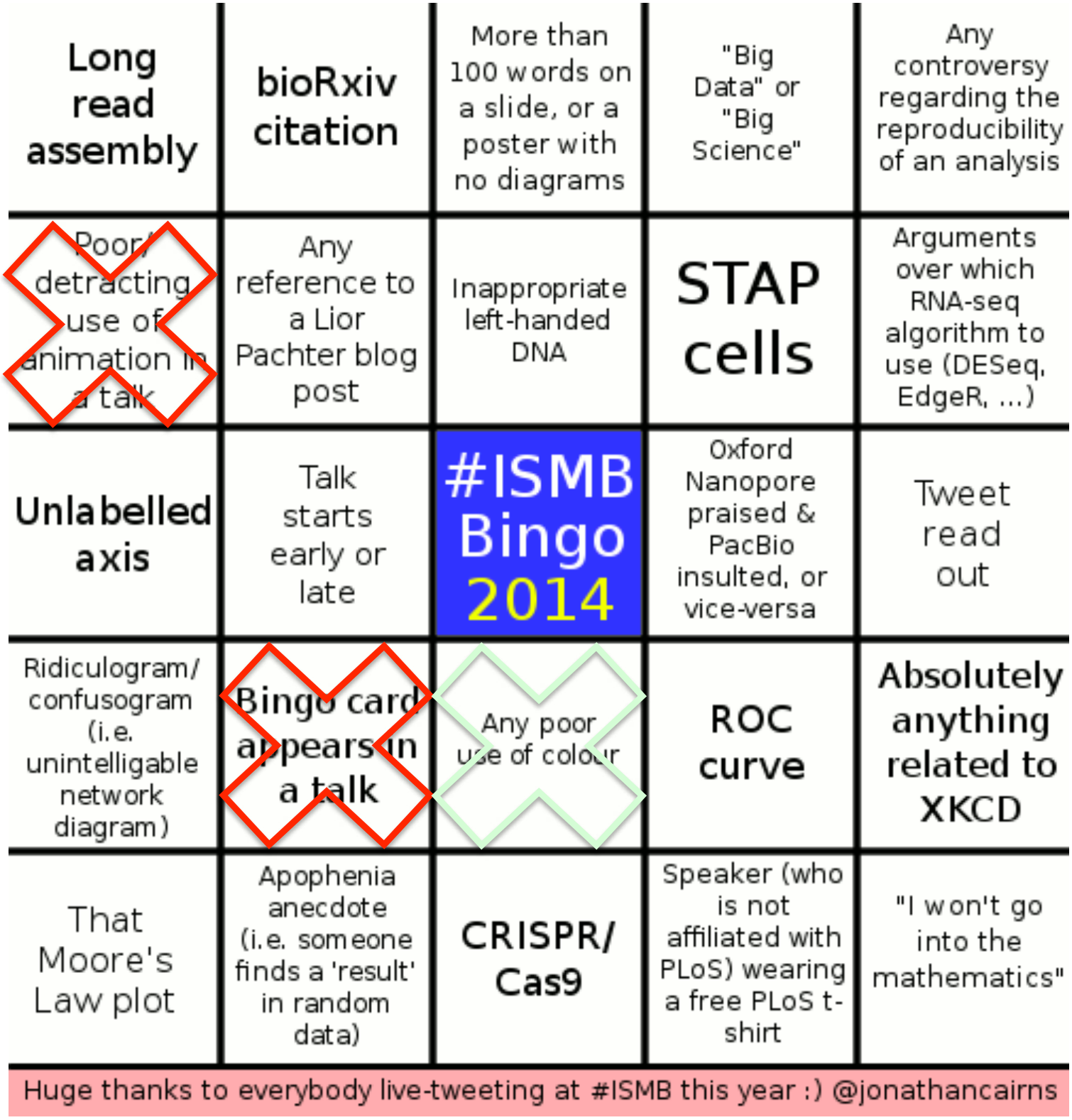
Funding

National Human Genome Research Institute

amazon web services™

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™


Long read assembly
bioRxiv citation
More than 100 words on a slide, or a poster with no diagrams
"Big Data" or "Big Science"
Any controversy regarding the reproducibility of an analysis
Poorly detracting use of animation in a talk
Any reference to a Lior Pachter blog post
Inappropriate left-handed DNA
STAP cells
Arguments over which RNA-seq algorithm to use (DESeq, EdgeR, ...)
Unlabelled axis
Talk starts early or late
#ISMB Bingo 2014
Oxford Nanopore praised & PacBio insulted, or vice-versa
Tweet read out
Ridiculogram/ confusogram (i.e. unintelligable network diagram)
Bingo card appears in a talk
Any poor use of colour
ROC curve
Absolutely anything related to XKCD
That Moore's Law plot
Apophenia anecdote (i.e. someone finds a 'result' in random data)
CRISPR/ Cas9
Speaker (who is not affiliated with PLoS) wearing a free PLoS t-shirt
"I won't go into the mathematics"
Huge thanks to everybody live-tweeting at #ISMB this year :) @jonathancairns

FIN