GENOMESPACE: An Environment for Frictionless Bioinformatics

www.genomespace.org

Michael Reich, Ted Liefeld, Marco Ocana, Dongkeun Jang, Jon Bistline, James Robinson, Peter Carr, Barbara Hill
Nathalie Pochet, Diego Borges-Rivera, Thorin Tabor, Helga Thorvaldsdóttir, Aviv Regev, Jill P. Mesirov

BROAD INSTITUTE

Features

GenomeSpace makes it easy for researchers to use the tools they already know to perform analyses and to find other tools that can help them extend their research into new areas. **Registration is free and includes 20GB of cloud based storage.**

GenomeSpace features include:

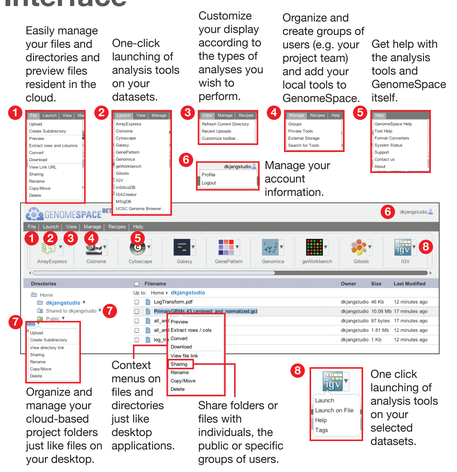
Seamless transfer of data between tools
GenomeSpace automatically converts file formats, removing the need to write scripts and "glue" code.

Easy import of data from public repositories
Users can transfer data directly from Web-based resources to their genomics tools without the need to download first.

Connect your own cloud storage accounts
Add your own Dropbox, Amazon or (coming soon) Google Drive accounts easily.

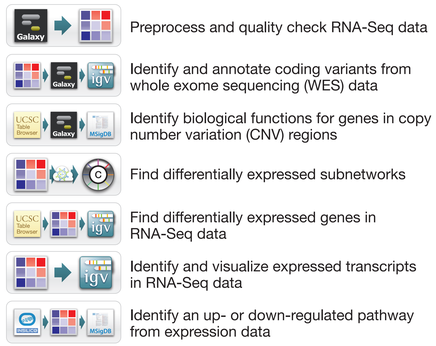
GenomeSpace is an NIH-funded project.

Interface



Recipes

A collection of "recipes" provides quick guides to accomplishing tasks using the GenomeSpace tools:

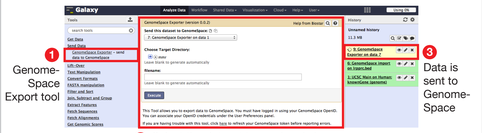


Using Galaxy with GenomeSpace

Galaxy users can send data easily between Galaxy and their GenomeSpace cloud storage:

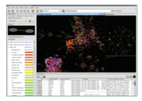
Importing Data from GenomeSpace to Galaxy

Exporting Data from Galaxy to GenomeSpace



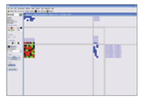
GenomeSpace Enabled Tools

Galaxy (Penn State University)
Scientific workflow and data and analysis platform providing a large number of sequence and genome analysis tools.

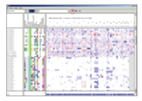


Cytoscape (UCSD)
Visualize and analyze molecular interaction networks and biological pathways

GenePattern (Broad Institute)
Analysis and workflow platform with hundreds of genomics tools for gene expression, sequence variation, proteomics, flow cytometry, etc.



Genomica (Weizmann Institute)
Analysis and visualization tool for integration of gene expression, DNA sequence, and gene and experiment annotation information.



IGV (Broad Institute)
High-performance visualization tool for interactive exploration of large, integrated genomic datasets.

UCSC Browser (UCSC)
Convenient access to the underlying UCSC browser database, including the reference sequence and working draft assemblies for a large collection of genomes.

Cistrome (Dana-Farber)
ChIP-chip- and seq-tools for peak calling and correlation, genome feature association, gene expression analyses, and motif discovery.

Gitools (PRBB)
Framework for analysis and visualization of genomic data using interactive heatmaps that also allows data to be imported from various sources.

InSilico DB (University of Brussels)
Powerful and intuitive interface to a large repository of gene expression datasets, including the contents of Gene Expression Omnibus (GEO).

ArrayExpress (EMBL-EBI)
Database of functional genomics (microarray and HTS) data

geWorkbench (Columbia University)
Integrated suite of tools for the analysis and visualization of gene expression, sequence, protein structure, and systems biology.

Synapse (Sage Bionetworks)
Platform designed to facilitate collaboration within and among scientific teams, access to large-scale genomics data sets, and integration with analysis tools and programming environments.

ISAcreator (U. of Oxford)
Metadata tracking tools to manage life science, environmental, and biomedical experiments.

MSigDB (The Broad Institute)
Online tools for a large curated collection of annotated gene sets.

Reactome (OICR) - Available soon
Reactome is a curated database of pathways and reactions (pathway steps) in human biology.

GenomeSpace Enabled Portals

Project Achilles
is a systematic effort aimed at identifying and cataloging genetic vulnerabilities across hundreds of genomically characterized cancer cell lines.

The Multiple Myeloma Genomics Portal
provides access to a reference set of multiple myeloma data as well as selected published multiple myeloma datasets.

The Cancer Cell Line Encyclopedia
provides public access to genomic data, analysis and visualization for approximately 1000 cancer cell lines.

How You Can Participate

We are seeking genomic researchers, bioinformatics tool developers, and data repository providers who are interested in joining and expanding the GenomeSpace community. See www.genomespace.org

Finding transcription factor regulators of human hematopoiesis

This example GenomeSpace scenario reproduces part of the Differentation Map analysis from the Regev lab paper in Cell, Novershtern et al, 2010

GENOMESPACE in Action

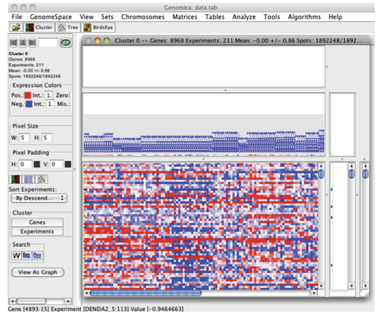
At each step, GenomeSpace performs all data conversions and transfers between tools.

1. User saves the expression data from the GO transcription factors to GenomeSpace.
2. User performs differential expression using the expression data loaded from GenomeSpace.
3. User loads the lineage-specific transcription factors generated in GenePattern to Genomica through GenomeSpace.
4. User uploads bed annotation tracks to Galaxy and IGV through GenomeSpace.

1 Genomica

Extract transcription factors

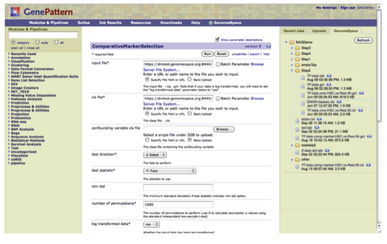
a. Load expression data containing 200 samples and 8000 genes
b. Load a gene set containing Gene Ontology (GO) transcription factors
c. Save the expression data from only the GO transcription factors to the GenomeSpace Data Manager.



2 GenePattern

Compute differentially expressed transcription factors

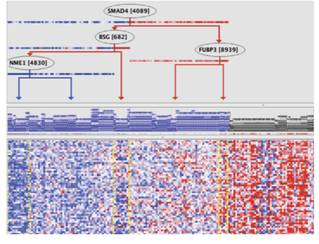
a. Perform differential expression analysis to determine genes that significantly distinguish human embryonic stem cells (hESCs) versus differentiated cells.



3 Genomica

Identify module networks

a. Compute module networks to determine coexpressed "modules" of genes within the original expression dataset.
b. Load the lineage-specific transcription factors generated by GenePattern
c. Use these two datasets to generate a list of potential regulators



4 Galaxy

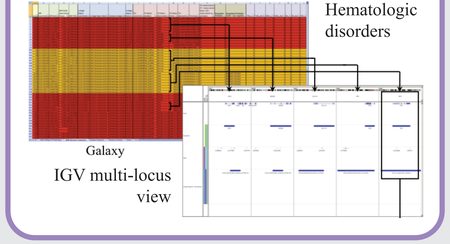
Compute overlaps

a. Upload annotation tracks for the genomic locations of the regulators, a set of previously published SNPs and a set of linkage regions from a genome-wide association study.
b. Run an overlap analysis to determine the intersection of putative regulators, SNPs, and linkage regions

5 IGV

Visualize data

a. Load annotation tracks for the 3 types of data in step 4 into IGV
b. View the concordance between the locations of the analytically identified potential regulators and the previously published SNPs and linkage regions



UC SANTA CRUZ · BROAD INSTITUTE · מכון ויצמן למדע · UCSD · PENNSTATE · STANFORD UNIVERSITY · National Human Genome Research Institute