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
Judith McLaughlin, Nathalie Pochet, Diego Borges-Rivera, Thorin Tabor, Helga Thorvaldsdóttir, Aviv Regev, Jill P. Mesirov

BROAD INSTITUTE

Background

Genomic research increasingly involves the generation and analysis of data across multiple modalities, e.g. sequence variation, gene expression, epigenetics, proteomics. These efforts are limited however by the difficulty of analyzing and integrating results from these multiple modalities. Each mode has its own tools, and the tools are seldom designed to work together.

To address these challenges, we have developed GenomeSpace, a lightweight, cloud-based infrastructure to allow genomics tools to share data seamlessly. GenomeSpace aims to knock down the barriers between tools, freeing researchers to perform analyses and investigate hypotheses that previously were too difficult to consider.

GENOMESPACE = Connect + Collaborate + Combine

Features

GenomeSpace makes it easy for biologists 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. GenomeSpace features include:

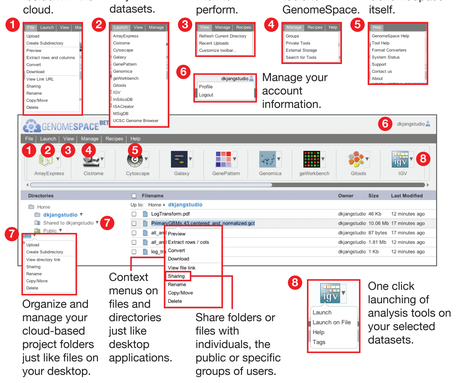
Seamless transfer of data between tools
GenomeSpace automatically converts file formats, removing the need to write scripts and "glue" code.

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.

Interface

1. Easily manage your files and directories and preview files resident in the cloud.
2. One-click launching of analysis tools on your datasets.
3. Customize your display according to the types of analyses you wish to perform.
4. Organize and create groups of users (e.g. your project team) and add your local tools to GenomeSpace.
5. Get help with the analysis tools and GenomeSpace itself.
6. Manage your account information.
7. Organize and manage your cloud-based project folders just like files on your desktop.
   Context menus on files and directories just like desktop applications.
   Share folders or files with individuals, the public or specific groups of users.
8. One click launching of analysis tools on your selected datasets.



Recipes

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

- Find differentially expressed subnetworks
- Find differentially expressed genes in RNA-Seq data
- Preprocess and quality check RNA-Seq data
- Identify and visualize expressed transcripts in RNA-Seq data
- Identify and annotate coding variants from whole exome sequencing (WES) data
- Identify biological functions for genes in copy number variation (CNV) regions
- Identify an up- or down-regulated pathway from expression data

Driving Biological Projects

GenomeSpace development is done in collaboration with two Driving Biological Projects (DBPs), which provide scientific direction as well as a collection of target research scenarios and analytic workflows.

1. **Dissection of regulatory networks in cancer stem cells by comparative network analysis with embryonic stem cells.** Regulatory networks of embryonic, induced pluripotent, and induced cancer stem cells are compared to find key differing regulatory networks. (Chan Lab, Stanford)
2. **Functional characterization of lincRNAs in mammalian genomes** by integrating epigenomic, transcription, RNA sequencing, RNAi, WGAS, and other modalities. (Regev Lab, Broad)

Seed Tools

- Cytoscape (UCSD)
- Galaxy (Penn State University)
- GenePattern (Broad Institute)
- Genomica (Weizmann Institute)
- IGV (Broad Institute)
- UCSC Browser (UCSC)

New Tools

- Cistrome (Dana-Farber)
- InSilico DB (University of Brussels)
- ArrayExpress (EMBL-EBI)
- geWorkbench (Columbia University)
- Synapse - *Available soon* (Sage Bionetworks)
- ISAcreator (U. of Oxford)
- Reactome - *Available soon* (Ontario Institute for Cancer Research)
- MSigDB (The Broad Institute)

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

GENOMESPACE in Action

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

Finding transcription factor regulators of human hematopoiesis

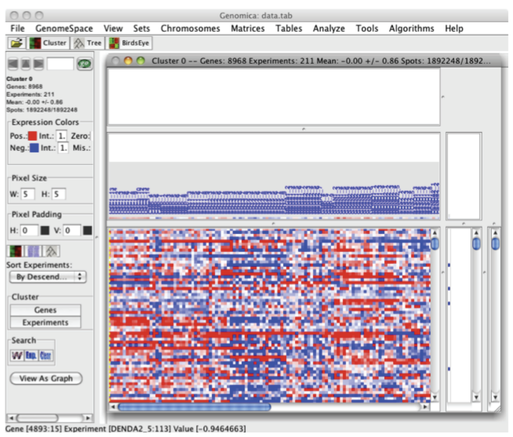
This example GenomeSpace scenario reproduces part of the DMap analysis from the Regev lab paper in Cell, Novershtern et al, 2010

1. **User saves the expression data from the GO transcription factors to 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.

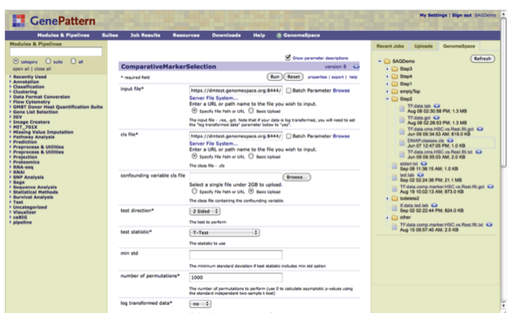


2. **User performs differential expression using the expression data loaded from GenomeSpace.**

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.



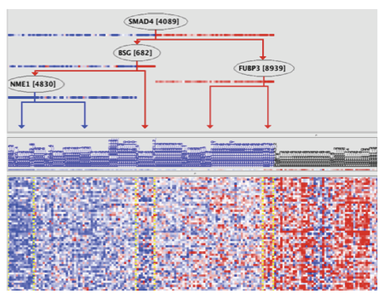
3. **User loads the lineage-specific transcription factors generated in GenePattern to Genomica through GenomeSpace.**

4. **User uploads bed annotation tracks to Galaxy and IGV through GenomeSpace**

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

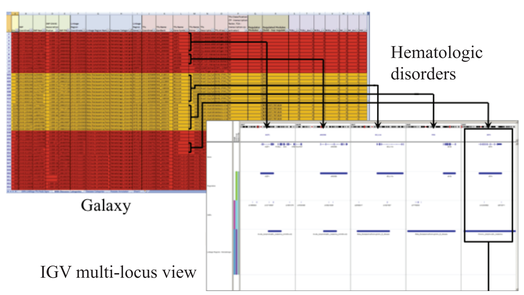
Compute overlaps

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 WEIZMANN INSTITUTE OF SCIENCE · UCSD · PENNSTATE · STANFORD UNIVERSITY · National Human Genome Research Institute