GENOMESPACE: An Environment for Frictionless Bioinformatics

www.genomespace.org

Michael Reich, Ted Liefeld, Jared Nedzel, Marco Ocana, Eliot Polk, Dongkeun Jang
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 are developing 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.

Connect

Collaborate

Combine

UC SANTA CRUZ · BROAD INSTITUTE · Weizmann Institute of Science · UCSD · PENNSTATE · STANFORD UNIVERSITY · National Human Genome Research Institute

Features

GenomeSpace will make 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:

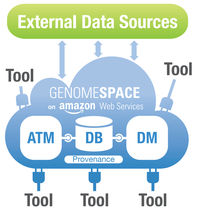
Seamless transfer of data between tools
GenomeSpace automatically converts file formats, removing the need to write scripts and "glue" code.

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.

Cloud based storage of data
After uploading your data to the cloud, it is accessible to all GenomeSpace tools.

Technology

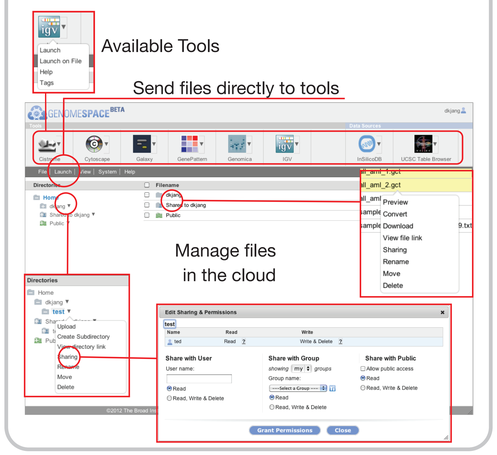
GenomeSpace allows users to store data in the Amazon Cloud, where it is accessible to all tools in the GenomeSpace environment. An easy-to-use Web services interface allows developers of bioinformatics tools and Web resources to add their software to GenomeSpace quickly and seamlessly.



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.geneomespace.org

Interface



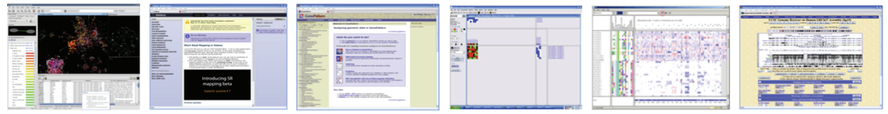
Driving Biological Projects

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

Six well-established genomic software packages form the "seed tools" that will make the GenomeSpace environment a comprehensive resource for analysis of heterogeneous genomic data types. GenomeSpace is being developed in collaboration with the development teams of these tools.



Cytoscape (UCSD) · Galaxy (Penn State University) · GenePattern (Broad Institute) · Genomica (Weitzmann Institute) · IGV (Broad Institute) · UCSC Browser (UCSC)

New Tools

Cistrome (Dana-Farber Cancer Institute) · InSilico DB (University of Brussels)

GENOMESPACE in Action

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

Finding transcription factor regulators of human hematopoiesis

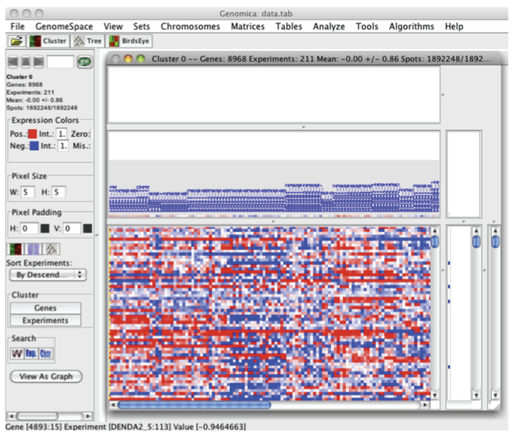
The first scenario supported by the GenomeSpace is to reproduce part of the DMap analysis from the Regev lab paper in Cell, Novershtern et al, 2010

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

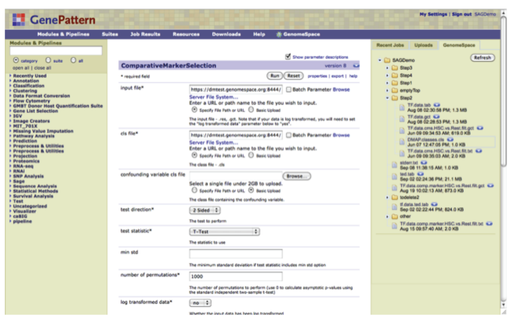
a. Load expression data containing 200 samples and 8000 genes
b. Load a gene set containing Gene Ontology (GO) transcription factors
c. Save the expression data from only the GO transcription factors to the GenomeSpace Data Manager.



2 GenePattern

Compute differentially expressed transcription factors

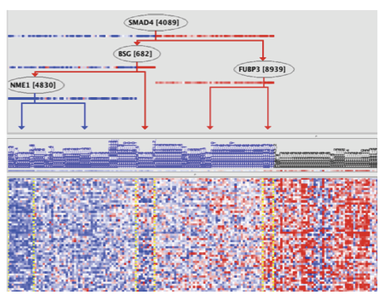
a. Perform differential expression analysis to determine genes that significantly distinguish human embryonic stem cells (hESCs) versus differentiated cells.



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 Galaxy
b. View the concordance between the locations of the analytically identified potential regulators and the previously published SNPs and linkage regions