

Genomics Portals: Integrative Data Analysis Tool

Mukta Phatak¹, Kaustubh Shinde¹, Johannes M. Freudenberg¹, Jing Chen¹, Li Q¹, Vineet Joshi¹, Hu Zhen¹, Krishnendu Gosh¹, Mario Medvedovic¹

Key words: Integrative data analysis, microarray data, epigenomics data, pathways, genomics database.

1 Introduction

A large amount of genomic experimental data generated by modern high-throughput technologies is available through various public repositories. Our knowledge about molecular interaction networks, functional biological pathways, gene regulatory modules, epigenomics data is rapidly expanding. Jointly these two sources of information hold a tremendous potential for gaining new insights into function of living systems. Genomics Portals is a web-based integrative platform for the analysis of diverse genomics data. It provides the interface for analyzing and interpreting one's own experimental data and the platform for mining of a large number of diverse, curated genomics datasets stored in the local databases. The server is also leveraging the existing web servers for functional analysis of gene lists (e.g. DAVID, L2L) and graphically integrating analytical results with pathways information (KEGG and DAVID).

2 Data Repository

Initial contribution of this work aimed towards curating and organizing vast amount of data in an efficient way. Genomics Portals comprise of two main components.

- 1) Gene lists: Back-end databases currently contain 13,381 pathways and other functionally coherent gene lists (KEGG, GO, L2L and in-house created gene lists). It also provides an option to upload custom gene list of interest.
- 2) Experimental assays: Presently we have 13,634 genome-scale vectors of measurements (more than 320 million data points) produced by various experimental assays (expression microarray, CGH microarrays, ChIP-chip and ChIP-seq) or computationally constructed scores (transcription factor binding scores and microRNA target scores). Expression profiles include, but not limited to, the data from breast cancer, prostate cancer, developmental models, toxicogenomics arrays etc.

In designing Genomics Portals we sought to strike a balance between the key limiting factors such as complexity of query interfaces, computational complexity of the analyses performed on the data and usefulness of the results produced. User friendly interface has been designed to query, navigate and analyze data.

3 Integrative Data Analysis

The portal aids investigation of genes of interest across different data sets. For example, pre-clustering of experimental assays is provided in an interactive TreeView session along with functional analysis of those clusters. One can select interesting cluster of mouse genes from one of the developmental expression profiles and can query human tissue profiling set using those genes.

Users can upload their own gene list of interest or construct a list using existing gene lists in the portal. For instance, we can choose as a input query list all genes in GO categories comprising of "stem cell" keyword in the description field. We can proceed to analyze one of the breast cancer microarray datasets for the ability of these genes to separate between different tumor grades and graphically integrate the results of the analysis with all KEGG pathways containing any of these genes. Corresponding pathways

¹ [Dept. of Environmental Health, Univ. of Cincinnati, Cincinnati, OH 45267, USA. E-mail: phatakma@email.uc.edu](mailto:phatakma@email.uc.edu)

will highlight statistically significant genes from the query list. Statistical analysis summary tables as well as hierarchical clustering heatmaps are available to browse and download. Similar analysis can be repeated for the same query list but using other datasets such as computationally predicted targets of different transcription factor, comparative genomic hybridization data for a related breast cancer datasets, genome-scale DNA methylation datasets for different tissue types etc. to gain more biological insight.

A typical analysis (outlined in the figure 1), starting by constructing a gene list of interest, querying one of the databases with genome-scale data, producing clustering as well as statistical analysis summaries and graphically integrating them within all affected KEGG pathways can take less than a minute.

4 Discussion

As the research interest in high throughput increases, “Genomics portals” brings a useful resource to the research community and as it evolves can bring broader perspective to the analysis of vast genomics data. The uniqueness of the portal lies in its ability to conduct integrative analysis that effectively connects heterogeneous pieces of information for the same gene list of interest under different conditions. To the best of our knowledge, no other public online resource provides similar level of functionality for publicly available experimental data. This composite analysis will help researchers detect novel information about their data and enable building new hypotheses.

Genomics Portal is open access and is available at <http://www.GenomicsPortals.org>

5 Figures

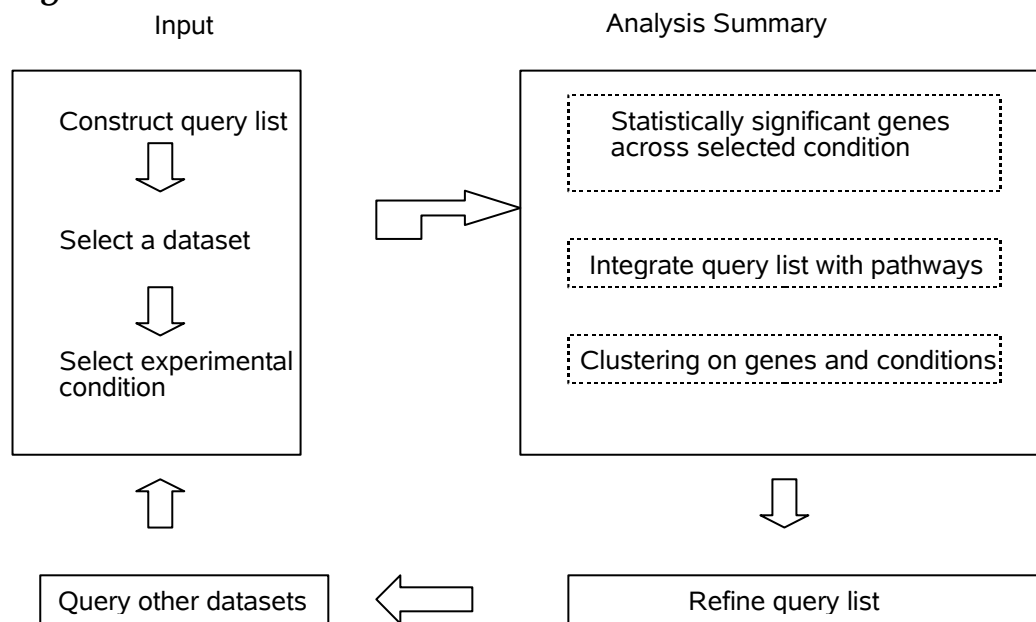


Figure 1: Usage flow

6 Acknowledgments

This research was supported by grants from the National Human Genome Research Institute (R01 HG003749), National Library of Medicine (R21 LM009662) and NIEHS Center for Environmental Genetics grant (P30 ES06096).