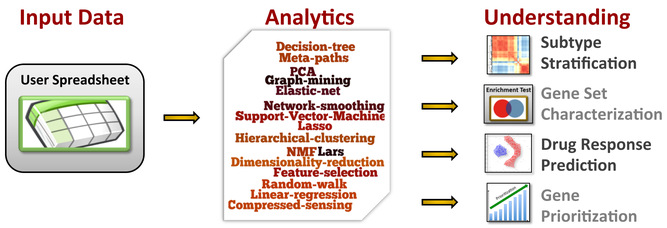
KnowEnG Cloud-based Scalable Analytics Suite

C. Blatti[1], M. Berry[2], L. Gatzke[2], A. Emad[1], N. Sobh[1], C. Bushell[1,2], S. Sinha[1,3]
[1]KnowEnG Center, [2]Applied Research Institute, [3]Department of Computer Science,
University of Illinois, Urbana-Champaign, IL

KNOWENG BIG DATA TO KNOWLEDGE CENTER OF EXCELLENCE

Analysis of Genomic Data

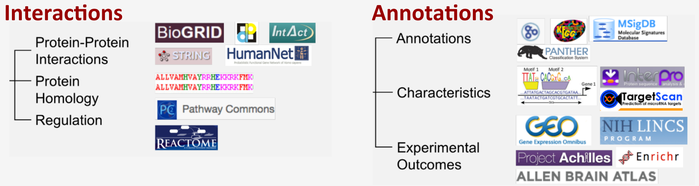
Researcher has experimental genomic profiles of multiple samples stored in a **spreadsheet**. Examples:
- genes x tumor samples somatic mutation matrix
- genes x cell lines basal transcription from RNA-seq



The user wants to apply **machine learning** and **graph mining** analysis to their spreadsheet in order to gain understanding of their experimental results.

Using Prior Knowledge

Each analysis workflow is designed to incorporate methods that take advantage of large collections of **public biological knowledge**. This knowledge includes many and varied interactions between or characterized annotations of genes and proteins across many species.

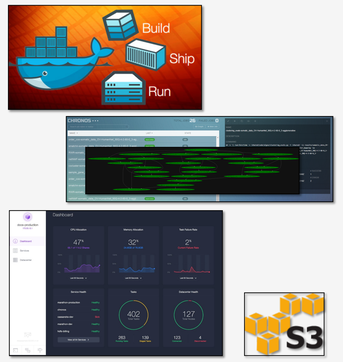


This collection of biological information is called the **Knowledge Network**. It is standardized for computation and made available by KnowEnG through databases and cloud storage.

On Scalable Cloud Infrastructure

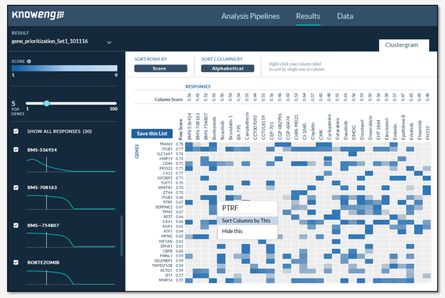
KnowEnG uses distributed systems, algorithms, and workflows that can scale to handle large-scale bioinformatics analysis on the increasing size and diversity of user data and community knowledge.

We deploy our analysis pipelines:
- As a series of **Docker containers** for each analysis task
- Whose execution is orchestrated by the **Chronos** job scheduling framework
- On a compute cluster managed by **Apache Mesos**
- That syncs user and community data through **AWS S3**



Analysis Workflows in KnowEnG

Gene Prioritization



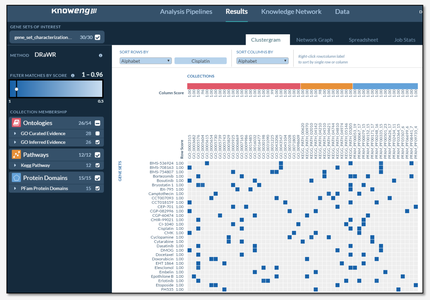
Goal: Identify genes whose mRNA expression levels explain the variation of drug sensitivity in different cell lines/individuals

Knowledge Network Hypothesis: Interesting genes for follow up analysis may show correlations of their neighbors' expression values with the drug sensitivity measurements

Approach: Robust network ranking of network smoothed expression profiles. Based on:

Knowledge-Guided Prioritization of Genes Determinant of Drug Resistance.
Emad A, Cairns J, Kalari K, Wang L, Sinha S.

Gene Set Characterization



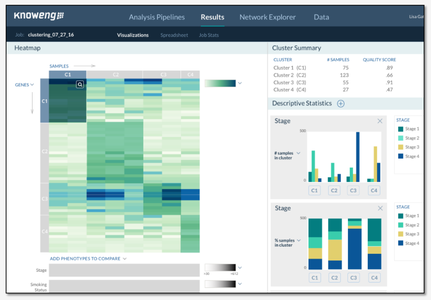
Goal: Find associations between the researchers novel gene sets and previously known biological annotations in order to provide understanding and hypothesis

Knowledge Network Hypothesis: Extends associations to poorly/incompletely annotated domains by integrating multiple relationship types

Approach: Network ranking of public gene sets for specificity to the query. Based on:

Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks.
Blatti C[1], Sinha S[2].
Bioinformatics

Subtype Stratification



Goal: Identify cancer subtypes from somatic mutation genomic data that are predictive of clinical outcomes

Knowledge Network Hypothesis: Within a subtype, two tumors may not share the same somatic mutation, but may have somatic mutations affecting the same pathways and protein interaction networks

Approach: Robust clustering of network smoothed profiles. Based on:

Network-based stratification of tumor mutations
Matan Hofree[1], John P Shen[2], Hannah Carter[2], Andrew Gross[3] & Trey Ideker[1–3]
nature methods

Upcoming Features

New Workflows
- ***Text Mining*** – Find genes most specifically related to different disease terminology
- ***Phenotype Prediction*** – Create model that predicts phenotypic outcomes from genomic data
- ***Gene Regulatory Networks*** – Model interactions between transcripts and transcription factors

Integration with Other Clouds
- Import user spreadsheets directly from other cloud-based datasets like TCGA, LINCS
- Package analysis workflows for other cloud computation engines like Seven Bridges Cancer Genomics Cloud

Interface Improvements
- Automatic mapping of gene/protein identifiers
- Interactive network-based visualizations to view user and public data
- Save and export analysis results (including support for Bagit data format and publication in BDDS Data Repository)

Acknowledgement:
"This research was supported by grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov)."
NIH National Institutes of Health