Highlight 1: Cloud-based Pipelines for Knowledge-guided Analysis of Genomics Data. We have created a Cloud-based infrastructure called KnowEnG (‘Knowledge Engine for Genomics’) for knowledge-guided analysis of genomics data. The user uploads their data in the form of a spreadsheet to the KnowEnG interface, and the system performs powerful data mining and machine learning tasks on those data. The unique part of such analysis is that it is carried out while making intelligent use of prior knowledge in the public domain. Such prior knowledge is represented in the form of a massive heterogeneous network called the Knowledge Network, which aggregates information from nearly 100 externally curated databases. The user may choose from several analysis pipelines (Fig. 1) to deploy on their data, including:

1. **Gene prioritization.** Identify genes most likely to be associated with a phenotype.
2. **Phenotype prediction.** Train a model to predict a numeric phenotype value from an omics profile.
3. **Sample clustering.** Find related groups within a collection of omics profiles. This may be used for patient stratification from their transcriptomics or somatic mutation profile.
4. **Gene set characterization.** Identify shared properties in a previously identified set of genes.

Each pipeline is a complex workflow involving one or more algorithms for data processing and normalization, application of the core machine learning or statistical algorithm, as well as post-processing and visualization. We are working towards integrating the KnowEnG system with other major Cloud-based data repositories such as TCGA and LINCS as part of a bigger ecosystem under the Commons umbrella (Fig. 2).

[Fig 1. KnowEnG pipelines interface. Fig 2. KnowEnG as part of the future Commons.]

Highlight 2: Advanced analytics for cancer pharmacogenomics. We have used KnowEnG analysis pipelines to rank and prioritize pathways, regulatory proteins called transcription factors (TFs), and genes associated with individual variation in drug response. Three novel algorithms have been developed to achieve this. The NetPath algorithm employs a network-based dimensionality reduction technique called Diffusion Component Analysis (DCA) to identify pathways that are most closely related to genes whose expression levels are correlated with drug response. The ProGeni algorithm employs Random Walks with Restarts (RWR) to rank genes by their association with drug response variation. The pGENMi algorithm uses Probabilistic Graphical Models (PGM) to integrate genotype (SNP), DNA methylation, gene expression, and phenotype measurements on a panel of cell lines, along with prior knowledge of the regulatory genome (ENCODE project), to identify transcription factors likely to influence drug sensitivity (Fig. 3). We have experimentally validated 23 predictions of genes associated with specific drugs.

[Fig 3. Integration of omics data.]

Highlight 3: Literature mining for disease-specific proteins. Typical text mining tools take substantial human efforts for manual data curation and extraction of structure from text data. In contrast, we have been developing a data-driven, semi-supervised text mining paradigm to mine a massive collection of biomedical texts. To ensure such effort will benefit biomedical research, we have been collaborating with the HeartBD2K Center at UCLA to conduct biomedical literature mining for disease-specific proteins. We analyzed hundreds of thousands of research abstracts in PubMed related to cardiovascular diseases in the last 20 years, comparatively ranking 250 given proteins with respect to each of six subcategories of cardiovascular disease. The results reveal new insights on disease-specific proteins. We are further proceeding to uncover distinctive dense protein subnetworks for the subcategories of heart diseases. We are also developing scalable literature retrieval functions for effectively finding (ranking) research papers given a set of biological entities (such as a set of genes). These biological text mining functions will be built into the KnowEnG engine and made accessible to the general biomedical research community.

The codebase for the KnowEnG tools is available at [https://github.com/KnowEnG](https://github.com/KnowEnG).