Interactive omics data exploration in epidemiological studies

Bjørn Fjukstad[1], Karina Standahl Olsen[2], Mie Jareid[2], Eiliv Lund[2] and Lars Ailo Bongo[1]

1) Department of Computer Science, UiT - The Arctic University of Norway
2) Department of Community Medicine, UiT - The Arctic University of Norway

Interactive data exploration is the process of exploring datasets looking for relationships and patterns in the data. In scientific disciplines such as epidemiology, data exploration enables an agnostic analysis approach that may be used to find hypotheses and insights not envisioned when the study was designed. Such exploratory analyses require systems that provide interactive visualizations and statistical analyses of large-scale omics datasets.

We used the Norwegian Women and Cancer (NOWAC) cohort to investigate and develop a requirement analysis for omics data exploration systems for epidemiological cohorts. We used the requirement analysis to develop **Kvik**, a framework for developing omics data exploration applications.

Systems for interactive exploration of epidemiological cohorts require:

i) familiar interfaces and visual representations
ii) lightweight applications making it possible to explore omics data on commodity computers
iii) powerful statistics support for advanced analyses
iv) interfaces to online knowledgebases.

With Kvik we have developed multiple applications for exploring omics data from both cross-sectional and nested case-control study designs. Our experiences using interactive applications for exploring epidemiological data proved superior to traditional workflows.

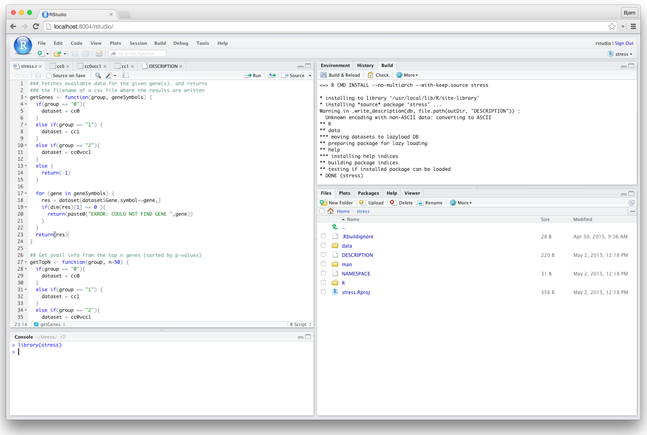

Figure 1: Researchers often write the data anlyses in programming languages such as R, making use of the wealth of available libraries. While omics datasets can fit on commodity computers, the sofisticated analyses require powerful machines to complete within seconds and not hours. In Kvik we placed computation on a fat server while researchers can explore the data on their own light computers with little computational resources.

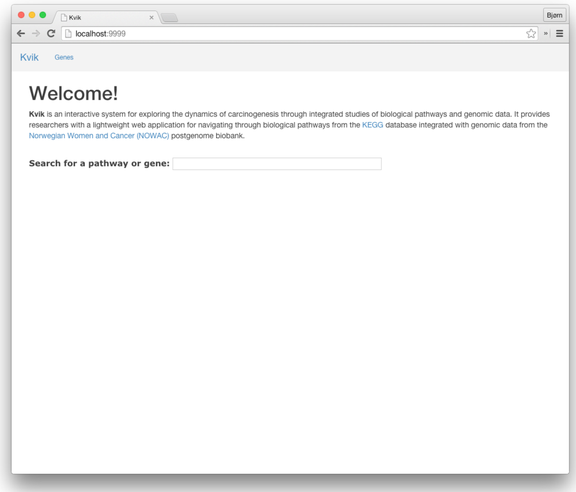

Figure 2: The welcome screen of one application for interactive exploration of omics data. This application runs analyses on the Stallo Supercomputer and presents the results in a interactive web application. The Kvik Stress Application allows users to explore data through targeted search for genes or pathways. Users can also start exploring the data from lists of differentially expressed genes from the dataset.

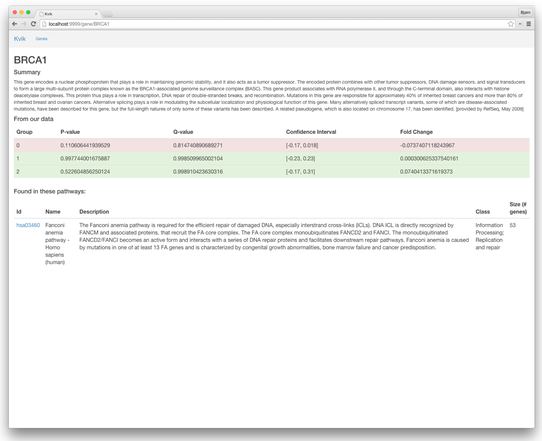

Figure 3: A user has searched for the BRCA1 gene. The search results contains both data from the data anlysis in addition to relevant background information about the gene and the pathway it is found in. By connecting to data sources such as KEGG (kegg.jp) or GeneCards (genecards.org)

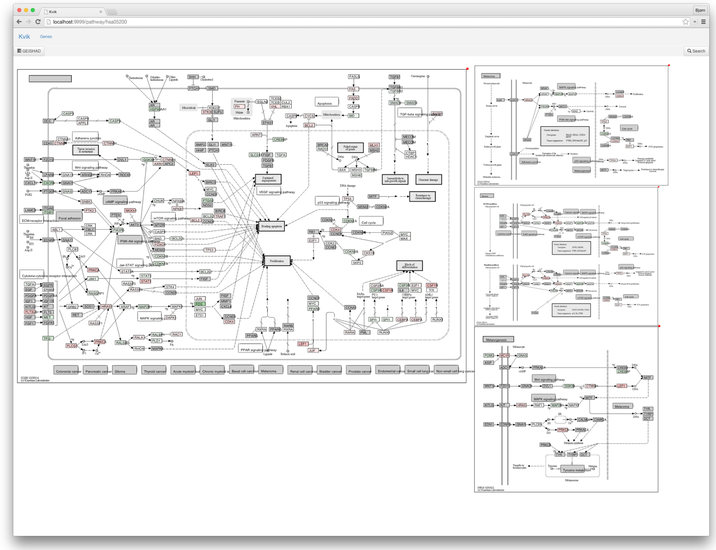

Figure 4: User exploring gene expression data in context of four different pathways. The application uses the familiar KEGG pathway layout to present the data.

github.com/fjukstad/kvik