Kvik: Interactive exploration of genomic data from the NOWAC postgenome biobank

Bjørn Fjukstad, Karina Standahl Olsen, Mie Jareid, Eiliv Lund and Lars Ailo Bongo

University of Tromsø – The Arctic University of Norway

{bjorn, larsab}@cs.uit.no

{karina.standahl.olsen, mie.jareid, eiliv.lund}@uit.no

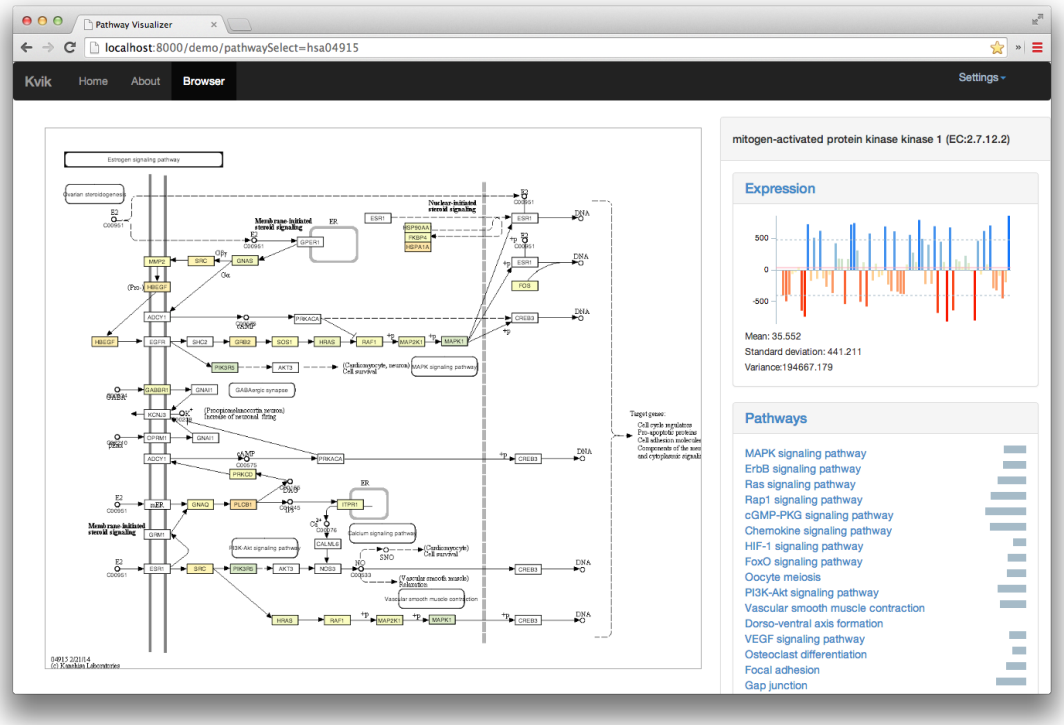

Figure 1: Kvik user interface. The Kvik Browser consists of a pathway visualization tool on the left and a gene information panel on the right.

Abstract

We have developed Kvik, a system for interactive exploration of genomic data from the Norwegian Women and Cancer (NOWAC) postgenome biobank. The goal of the NOWAC study is to understand the dynamics of carcinogenesis through multi-level functional analyses of transcriptomics and epigenetics using blood and tissue samples. Kvik provides a tool for exploring gene expression data, incorporating both statistical analysis and interactive visualizations in a single system. The tool is open-sourced at `github.com/fjukstad/kvik`.

This paper was presented at the NIK-2014 conference; see nik.no.

1 Introduction

The recent advances in omics technologies, such as Next-generation sequencing (NGS) machines, has the potential of producing data that provide views of biological processes at different resolutions and conditions, enabling novel insights into the development of cancer (carcinogenesis). In addition, the cost of NGS machines and analyses have become low enough that it is practical to purchase and operate these machines in individual research labs or in clinical care. However, these machines produce large amounts of data [1]. Downstream analysis and interpretation of the produced data is therefore a major challenge [2].

A promising approach for novel understanding of carcinogenesis is through explorative data analysis of biobanks such as the Norwegian Women and Cancer (NOWAC) postgenome biobank. NOWAC is the realization of a systems epidemiology research project designed to identify relationships between lifestyle and the risk of cancer. The project started the data collection in 1991, and by 2006 the study contained questionnaire information from 170,000 women. Since the collection of biological data started in 1998, the NOWAC biobank has grown to over 60,000 blood samples and 800 biopsies that have been, or will be, analyzed using whole-genome gene expression analysis tools. In the NOWAC study, researchers have the opportunity for interactive studies of the human genome, transcriptome, proteome and metabolome, known in total as the globolome [3]. Together with lifestyle exposure information through questionnaires, the globolomic design has the opportunity for merging epidemiological research with analysis of large-scale data from high-throughput platforms.

However, to enable researchers to transform the large quantities of data in the NOWAC biobank into novel insights about carcinogenesis, new data exploration systems and tools must be developed. In particular, there is a need for visualization tools that integrate data from multiple biological levels (e.g. from protein molecules up to complete organisms), link these to cancer meta-databases, and present these in an integrated interactive system. Due to the novelty of the design of the NOWAC biobank and the corresponding analysis methods, no existing tool exist for such explorative analyses.

We present Kvik, a novel approach for exploring large-scale omics data and analysis of biological pathways. Kvik provides the integration of gene expression data and biological pathways in a single browser, providing an interactive visualization tool. Researchers can view and analyze pathway maps combined with research data such as high-throughput DNA sequencing data. Kvik combines gene expression data and state of the art pathway maps, and together with an information panel it provides detailed information about single genes as well as the underlying experimental data and design.

Kvik has the three-tier architecture commonly used for web-scale data exploration tools such as the Google search engine [4]. It provides a portable web application for researchers to browse pathways and data collected in the NOWAC study, and powerful computational backend for performing advanced statistical analyses. This separation allows Kvik to provide secure storage of genomic data, and researchers to perform exploration on ligtweight clients.While we integrated gene expression data from the NOWAC biobank with pathway maps, Kvik is a general purpose tool that can integrate genomic data from any study.

2 Biological Background

Carcinogenesis is the development of cancer in an organism. Cancers are the results of gene mutations that can occur in any cell anytime. These mutations may cause the cells to grow uncontrollably, resulting in tumors. Cancer is primarily a disease of old age, due to accumulation of mutations over years.

Some inherited mutations may increase the risk of cancer considerably, for example the well known link between inherited mutations in genes BRCA1 and BRCA2 and risk of breast cancer [5]. To understand carcinogenesis, researchers must often study changes in multiple genes that possibly participate in different biological processes. In addition, environmental factors such as sunshine, smoking or lack of exercise can impact the risk of cancer. It is therefore necessary to also take into account the epigenome to understand carcinogenesis.

Studying the changes in a number of genes and how this affects the processes they participate in is a challenging task. Not just because of the number of genes and processes, but also the fact that researchers may look for small changes in multiple genes, not necessary big changes in single genes. Another challenge is the time aspect of carcinogenesis, how are the genes expressed prior to diagnosis of cancer? Understanding this process requires the monitoring of genes over time series spanning decades. Studying gene-environment interactions requires novel epidemiological study designs that considers the complexity of the multistage carcinogenic process, latency and the changing lifestyle of the participants [6].

Humans have approximately 20,500 genes that determine how the body performs different biological processes. Genes are long sequences of Deoxyribonucleic acid (DNA), often represented as strings of four letters A, G, T and C. DNA sequencing is the process or determining the order of these letters. Additionally, while almost all cells in a human contain the 20,500 genes only a subset of these are active. This activity level is known as gene expression. DNA microarray technology or serial analysis of gene expression (SAGE) techniques are used to measure gene expression levels.

Collecting biological data used to be a time consuming and costly process, but with techniques such as NGS and DNA microarrays it is possible to retrieve the genetic information stored in DNA for only a fraction of the cost and time. The sequencing cost for a megabase (a million letters from a DNA strand) has already out-paced Moore's Law,[1] making it feasible to collect more genetic data. This wealth of data brings the potential to develop more accurate treatment and diagnosis methods. Previously it could have been possible to study genomic data in basic excel spreadsheets, but to explore the continuously growing datasets researchers must embrace high-performance computation systems for both analysis and storage of genomic data. Essential for such analyses are visualisation tools.

To understand the complex biological processes affected by carcinogenesis in an organism, researchers use abstract graphical representations of these, known as pathway maps. These human curated maps are essential for biologist to understand, for example how genes interact with each other, the assembly of new molecules and signal transmission within an organism. Together with pathway maps, DNA expression data gives researchers a more accurate depiction of the current processes in an organism. Researchers can study how gene expression levels affect different biological processes, e.g. how some genes cause over-production of a specific protein. With the globolomic design of the NOWAC study, it is even possible to study how lifestyle or environmental

[1] genome.gov/sequencingcosts

exposures affect different biological pathways. It is this exploration and sensemaking of epidemiological research data we want to support with Kvik.

3 Requirement Analysis

Kvik is based on collaboration with epidemiology researchers at the Institute of Community Medicine at the University of Tromsø. Together we performed a requirement analysis and identified six requirements for an interactive data exploration system for genomic data from the NOWAC postgenome biobank. In the following we discuss the requirements and how Kvik fulfills them.

Familiarity Kvik should provide visualizations familiar to the end-users. Kvik visualizes biological pathways using the widely-used Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway layout, achieving familiarity with contextual visual cues familiar to researchers. By integrating genomic data and relevant information from the KEGG databases, researchers are able to use a single tool for data exploration.

Interactivity Kvik should support explorative visual data analysis. This requires high-degree of interactivity. Kvik achieves interactivity by visualizing and integrating biological pathways with corresponding gene expression data and relevant information, using a single HTML5 view providing sub-second feedback to end-users.

Ease-of-use Kvik should provide an easy to use tool, both for the end users as well as the maintainers of the system. The exploration tool in Kvik runs in a web browser and does not require installing any third party plug-ins or applications. The backend system in Kvik consists of independent components that communicate through well-specified REST interfaces.

Scalability Kvik should scale to the upcoming datasets on the NOWAC study. In addition to data management and processing, the visualization tools should also be capable of visualizing large quantities of data. Kvik has a three-tier architecture where computational resources and display resources are separated. This makes it possible combine a backend data analysis and storage systems that run on a large cluster, with lightweight data exploration clients. Such an architecture has been shown to scale to peta-scale datasets [7].

Extensibility Researchers often collect data from a large number of sources, both online databases and their own datasets. In addition, researchers are continuously discovering novel methods to visualize or process data. It should therefore be easy to extend Kvik with new analysis methods, data sources, and data visualizations. Kvik incorporates multiple heterogeneous data sources into a single data engine, including online databases and data from local datasets. It uses R, a standard programming language for statistics, allowing simple implementation of new analysis methods. In addition, the modular architecture of Kvik makes it easy to utilize software and hardware improvements, such as data-intensive computing systems [7–10].

Security Kvik should provide an interface to access data from a secure storage facility, possibly behind restrictive firewalls. Kvik uses a designated data engine for storing expression data from the NOWAC study. Kvik will only expose visualizations to the researchers, making it impossible to view or download any of the raw data.

4 Kvik

Kvik is a system that combines powerful statistical analysis with interactive lightweight visualizations. Kvik separates computation and visualizations resources, allowing

researchers to explore genomic research data on lightweight clients while Kvik manages and performs computationally intensive statistical analysis on a backend separated from the users. Kvik provides exploration of gene expression profiles from the NOWAC postgenome biobank integrated with state of the part pathway maps from the widely used KEGG pathway database.

Kvik has a three-tiered architecture consisting of three independent layers: the *browser layer*, a lightweight web application for exploration of gene expression data and biological pathways; the *frontend layer* provididing static content such as HTML pages and stylesheets, as well as an interface to the data sources with dynamic content such as pathway maps and gene expression data; and the *backend layer* consisting of information about pathways and genes, as well as computional and storage resources to process the genomic data from the NOWAC biobank (Figure 2). Typically the Kvik Browser would run on a researchers' workstation or laptop, while the other components run in a cluster environment to provide storage and computation capabilities beyond desktop computers.

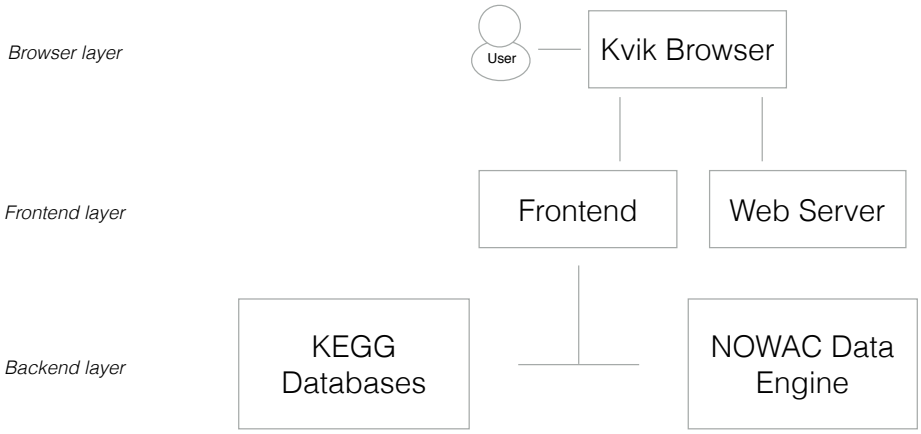

Figure 2: Kvik system architecture.

Kvik Browser

The Kvik Browser is the single point of interaction for users of Kvik. It provides a web application for interactively exploring gene expression data integrated in biological pathways. The Kvik Browser is a graphical tool that uses HTML5 technology. With a web application there are multiple advantages. First, users receive updates to the application and data by simply refreshing a webpage. Second, with HTML5 users don't need to install any third-party plugin or application, e.g. Java or Flash. Third, HTML5 is a technology that allows the application to run on mobile devices as well, making it possible for researchers to present or discuss results out of the office.

To generate the pathway visualizations, Kvik merges two different pathway representations from the KEGG databases. The KEGG databases provide static images representing different pathways, as well as a KEGG Markup Language (KGML) representation.[2] The KGML representation contains a description of the entities (nodes) and relations (edges) that make up the pathway map, providing information such as x and y coordinates for entities. Entities can be genes, proteins or compounds, and the relations are reactions between the entities, e.g. the activation of a gene. Unfortunately, the KGML file lacks edge routing information [11] and missing nodes. In addition,

[2]KGML is a Extensible Markup Language (XML)-like format following similar syntax, see `kegg.jp/kegg/xml`.

the manually curated pathway images bundle together multiple edges found in the KGML representation. Because of these imperfections, Kvik generates the pathway visualizations by rendering the static pathway image from KEGG and drawing nodes on top of this image. This allows Kvik to render gene expression data on top of the pathway maps while still providing the contextual information from the human curated pathway image. Kvik colors the nodes in the pathway maps according to gene expression values from the NOWAC Data Engine. It uses the Javascript libraries Cytoscape.js[3] and D3[4] to generate visualizations of biological pathways and gene expression data. Cytoscape.js is an open-source JavaScript graph library for analysis and visualization, that uses the HTML5 canvas element to render graphs, making it possible to visualize large networks.

Users can select different genes in the pathway maps to get a detailed view of the gene expression values for the entire NOWAC biobank, as well as other information from KEGG. The Kvik browser presents this information in a gene information panel adjacent to the pathway visualization.

Frontend and Web Server

The Frontend is the component that connects the Kvik Browser with the data sources in the backend layer. It is responsible for performing queries to either KEGG or the NOWAC Data Engine, and populating the Kvik Browser with data. The Web Server is responsible for hosting the Kvik Browser's static pages and stylesheets.

NOWAC Data Engine

Kvik uses a subset of the NOWAC biobank as the main source of gene expression data. At the time of writing the Kvik dataset consists of gene expression profiles collected at time of diagnosis for 77 women in addition to gene expression data for 77 controls (healty women). The dataset contains gene expression values for 9101 genes. For more information, refer to Lund et al., 2008 [3].

To provide a familiar framework for the researchers, we chose to implement the NOWAC Data Engine in the R[5] programming language. Among other features it provides support for the popular open-source genomic data library BioConductor.[6] The use of the R programming language allows researchers to contribute to the Kvik Data Engine with their own statistical analysis methods, and the addition of new methods from other bioinformatics libraries.

KEGG Databases

Kvik uses KEGG as the primary source of information. The KEGG database is frequently updated, and is freely available for academic users via their REST API.[7] There is also a FTP subscription available, but requires a licence. To avoid having to manage the database and keeping it updated, Kvik uses the KEGG REST API. Kvik hosts a local cache with requests to the KEGG database to reduce latency, but also to reduce the number of requests to the KEGG servers. Single requests to the KEGG servers located in Japan could result in response times in the order of seconds. However, using a local cache allows Kvik to

[3] `cytoscape.github.io/cytoscape.js`
[4] `d3js.org`
[5] `r-project.org`
[6] `bioconductor.org`
[7] `rest.kegg.jp`

perform requests to KEGG in the order of milliseconds. The cache writes every response from KEGG to disk, allowing the Frontend to retrieve results locally before contacting KEGG in case of a cache miss. Since all requests to KEGG come from the server-side of Kvik, two or more identical exploration sessions will only produce the network traffic of one session.

5 Evaluation

The main goal of the evaluation is to evaluate the design choices and implementation of Kvik. To do this, we evaluated Kvik using three metrics: i) latency, ii) resource utilization and iii) system scalability. In addition we conducted an informal evaluation of the system with regards to usability from an end-user's perspective.

The most important factors that impact the response time, resource usage and scalability of Kvik are:

Pathway Size The number of entities (genes, proteins or compounds) in a pathway affects the latency since the Kvik Browser requests information from KEGG for every entity. In addition to the information, the Kvik Browser retrieves gene expression values from the NOWAC Data Engine for every gene in the pathway.

Dataset Size Kvik uses a subset of the NOWAC biobank which contains 9102 expression values for 154 individuals. Increasing the dataset size, e.g. the number of patients, has an impact on the start-up time of the Data Engine. The execution time of simple summary statistics is not affected by the current dataset size. We expect the execution time to increase for the full NOWAC dataset, that is orders of magnitude larger than the subset used in this paper. In addition we plan to use more advanced statistical methods that are more compute-intensive.

KEGG Response Time Kvik requests meta-data for all genes and pathways from the KEGG database. Since Kvik uses the freely available REST API. KEGG query time has significant impacts on the end-to-end latency of the system.

To evaluate the latencies of the visualizations in Kvik, we measure the load times of different pathways. To evaluate the scalability of the NOWAC Data Engine, we measured both load time of the data as well as memory consumption. Finally, to evaluate the need for a local KEGG cache, we measured the visualization latency in Kvik without caching enabled.

Experimental Platform

Kvik was tested and evaluated using commodity hardware. All experiments were run on a 2013 Mac Mini with a 2.66 GHz Intel Core i7 processor, 16 GB 1600 MHz DDR3 SDRAM and a 1 TB Fusion Drive, connected to two 1920x1080 Acer P246HBD displays. The end-users and Kvik servers were located in Tromsø Norway, while the KEGG Database severs are in Japan.

Visualization latency

We chose four different pathways to measure the load times relative to pathway size (Table 1). These are all important for the analysis of NOWAC data. As of the time of writing, the KEGG pathway database holds 289 human pathways which have a mean size of 105 entities.

Table 1: Pathways used to evaluate Kvik.

Id	Name	Number of entities
hsa04630	Jak-Stat signaling pathway	35
hsa04915	Estrogen signaling pathway	74
hsa4151	PI3K-Akt signaling pathway	120
hsa05200	Pathways in cancer	267

Figure 3 shows the cumulative distribution of the measured latencies for the different pathways. The Kvik Browser completes 95% of the requests for the first three pathways within a second. For the large hsa05200 (Pathways in cancer) pathway, it completes 95% of the requests after about 2 seconds. As reported in Miller, 1968 [12], response times less than 1 second do not interrupt a user's line of thought. In Kvik all visualizations and events display a progress indicator to let users know that the system is performing some work. This will further increase the latency accepted by users. According to our users, Kvik generated pathway visualizations within a reasonable amount of time.

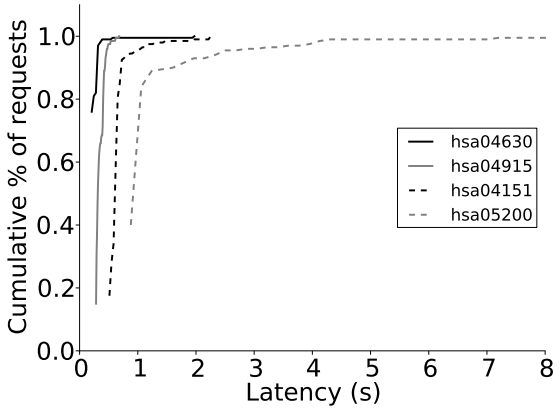

Figure 3: Cumulative distribution of the measured latency.

To reduce the latencies, the Kvik Browser could use client-side caching. With HTML5 it is possible to store up to 5MB of data client-side, allowing the Kvik Browser to cache both the pathway image as well as other information on the client.

Resource usage and scalability

To evaluate the NOWAC Data Engine it was necessary to generate test datasets. We added additional case-control pairs to increase the dataset size. While we expect the number of genes in humans to stay constant, the number of case-control pairs will increase as more blood samples or biopsies are analyzed. The test datasets contained random gene expression values. The NOWAC Data Engine was evaluated by measuring the initializing time and the memory usage after initialization. This includes reading in datasets and making them available to the Kvik Browser.

Table 2 shows the results from initializing the NOWAC Data Engine, with datasets up to 20x[8] the size of the dataset currently used by researchers in Kvik. The table shows how initialization time and memory usage grow for increasing dataset size. The initialization reads the dataset from disk and stores it in main memory. To load the original dataset

[8] 9102 genes and 1540 case-control pairs

consisting of 9102 genes and 77 case-control pairs takes on average 2.63 seconds and uses 353 MB of memory. As the results show both memory usage and load time appears to scale linearly. From the results it is apparent that the NOWAC Data Engine can hold the test datasets containing gene expression data for over 1500 case-control pairs in memory. For dataset sizes that do not fit in DRAM, it is necessary to either increase the DRAM size of the machine or use a distributed cache (such as BigTable/HBase [7] or Spark [9]). Today a 256GB DRAM server is reasonably cheap, but we expect a distributed cache to be necessary when data from additional instruments (methylation, NGS) are integrated. In addition, we believe that future advances in genomic research rely on high-performance systems for both storing and processing genomic data [13]. Recently, several research projects such as Google Genomics[9] or Illumina BaseSpace[10] show promising results in analyzing genomic data using cloud technologies.

Table 2: Runtime and memory usage for initializing the NOWAC Data Engine.

	Time		Memory	
Size	Mean	SD	Mean	SD
1x	2.63s	0.13s	353 MB	0.04 MB
2x	5.39s	0.26s	703 MB	0.06 MB
5x	13.94s	0.29s	1652 MB	0.10 MB
10x	28.06s	0.28s	3320 MB	0.19 MB
20x	58.18s	0.68s	6579 MB	0.85 MB

KEGG

To measure the impact of caching requests to KEGG, we ran the first set of experiments without caching enabled. From Table 3 we see the benefit in caching KEGG responses. The table shows the average load times for the four pathways, both with caching enabled and disabled. For every pathway measured there is over a 140x speedup in load times when caching KEGG responses. Visualizing the four experiment pathways resulted in 693 elements in the cache adding up to 5.7MB used disk space server side. We expect the cache to stay relatively small since many of the cached items (such as information about genes and compounds) are shared between pathways.

Table 3: Load times of KEGG requests.

Method	Caching Enabled	Caching Disabled
Load pathway hsa04630	0.20s	53.18s
Load pathway hsa04915	0.34s	59.63s
Load pathway hsa04151	0.62s	123.99s
Load pathway hsa05200	1.00s	142.30s

Development process

The development process of Kvik followed an iterative approach. We had frequent meetings with the end-users. We received continuous feedback about the usability and

[9] `developers.google.com/genomics`
[10] `basespace.illumina.com`

features provided by Kvik. The visualization approach used by Kvik is a result of these meetings.

We chose to generate pathway visualizations using the traditional layout of KEGG early in the development process. We explored other techniques, e.g. using a force-directed layout, but these maps were unusable by the researchers who were used to browsing the hand-drawn pathway maps in KEGG. In the early development stages, Kvik generated pathway visualizations based on only the KGML representations, but these quickly became too complex and difficult to navigate. When we visualized nodes on top of static KEGG images, the researchers were able to orient themselves in the traditional pathway maps, allowing them to start data exploration faster. Integrating a gene information panel within the same view was an important requirement from our collaborators. Removing the need to browse other databases for relevant information, researchers could browse and explore genes faster. This component took multiple iterations before the results were satisfactory to our collaborators. Throughout the project, we added features to Kvik using this iterative approach.

From the iterative development process, adding components to Kvik became simpler with every feature. It is evident that collaborating in a team with frequent meetings is helpful both for biologists to come up with improvements and new features, but also for computer scientists to understand the problems from the end users' perspective. Overall our collaborators were satisfied with Kvik, both in terms of usability but also in terms of the future opportunities with the project.

6 Related Work

To our knowledge, no existing system fulfill the requirements of a system for exploratory analysis of the NOWAC postgenome biobank. There are multiple online resources for visualizing biological pathways, like KEGG [14] or BioCarta [15], but they provide poor interaction support. These tools require users to switch between separate views when selecting genes or compounds in a pathway, making it difficult to keep the same line of thought when exploring the biological pathways. VisANT [16], VANTED [17] and KEGGViewer [18] are systems for interactively exploring biological pathways, but since these build on the KGML representations, these lack visual cues like cell walls that would make the visualizations familiar to biology experts. enRoute [19] and Entourage [20] both included the Caleydo framework [21] provide familiar visualizations and incorporates the possibility to visualize gene expression from multiple heterogeneous data sources. Nevertheless, the Caleydo framework is a standalone application that requires installation on researchers computers, failing the ease-of-use requirement. The Caleydo framework is not the only system failing the ease-of-use requirement. VisANT and Vanted are both dependent on users installing the Java Runtime Environment (JRE) in addition to the application, or a Java plug-in to run in the web browser. Pathway Projector [22] is a system that visualizes biological pathways without any installation. It allows users to browse biological pathways similar to viewing maps on Google Maps[11], but fails the security requirement since researchers must upload their dat to the Pathway Projector servers for visualization.

[11] `maps.google.com`

7 Future Work

Kvik fulfills our initial requirements and has proven useful. During development and deployment we have identified areas of improvement. While Kvik implements a simple exploration tool with a pathway browser and a gene information panel, our collaborators require the addition of more advanced analyses like Gene set enrichment analysis (GSEA) to perform more complex exploration of the NOWAC biobank. The addition of such analyses is possible through the extendible NOWAC Data Engine.

The first version of Kvik supports multiple users with access to the same dataset. Future versions of Kvik will support multiple users with access to different datasets and other private data. This feature will allow researchers to share results between each other, supporting collaboration on exploratory analyses.

Another feature requested by our collaborators is the ability to export a complete workflow, or session, similar to Galaxy or the syntax files in SPSS[12]. Users want to record the data exploration steps, e.g. the navigation through pathways and statistical analyses done. This will improve the reproducibility of the results found by using Kvik.

Currently, Kvik has only access to gene expression data at time of diagnosis. The NOWAC biobank consists of large quantities of other research data, such as gene expression and exposure through questionnaires that have been collected over decades. Our collaborators are developing new statistical analysis methods that integrate data from multiple times, levels and instruments. Since the NOWAC Data Engine is implemented in the R programming language, it has access to extensive bioinformatics libraries. Such libraries allow for easy integration of new statistical analysis methods and genomic data. The R programming language is suitable since our current datasets fit in DRAM. Future versions will require the use of high-performance systems such as Spark [9] to manage the large datasets.

The Kvik Browser is currently limited to visualizing one pathway at a time. We plan to incorporate visualization of multiple pathways in the same view, as in Entourage [20]. Navigating multiple pathways is often necessary to understand complex diseases such as cancer. Keeping them in the same view can speed up the exploration.

8 Conclusion

We have presented a novel tool for exploratory analyses of biological pathways and genomic data. Our approach allows researchers to explore large quantities of research data on lightweight clients. By separating visualization and computation resources, Kvik can manage and perform compute-intensive statistical analyses, while researchers can focus on making key insights and form new hypotheses by interactively exploring the NOWAC biobank.

Our approach to the exploration tool was based on an iterative collaboration between developers and end-users from the NOWAC research group. We provide a requirement analysis that targets the challenges of performing exploratory analysis of biological pathways and genomic data.

Through an evaluation of the exploration tasks and feedback from end users, we demonstrate that Kvik has the performance required of interactive exploration of genomic data and biological pathways. While Kvik targets exploration of biological pathways and gene expression data, we believe that the architecture is suitable for interactive exploration of data from other disciplines as well.

[12] `ibm.com/software/analytics/spss`

References

[1] S. D. Kahn, "On the Future of Genomic Data," *Science*, vol. 331, Feb. 2011.

[2] A. Sboner, X. Mu, D. Greenbaum, R. K. Auerbach, and M. B. Gerstein, "The real cost of sequencing: higher than you think!," *Genome Biology*, vol. 12, 2011.

[3] E. Lund, V. Dumeaux, T. Braaten, A. Hjartåker, D. Engeset, G. Skeie, and M. Kumle, "Cohort profile: the Norwegian women and cancer study—NOWAC—Kvinner og kreft," *International journal of epidemiology*, vol. 37, no. 1, 2008.

[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1, 1998.

[5] M.-C. King, J. H. Marks, J. B. Mandell, *et al.*, "Breast and ovarian cancer risks due to inherited mutations in brca1 and brca2," *Science*, vol. 302, no. 5645, 2003.

[6] E. Lund and V. Dumeaux, "Systems epidemiology in cancer," *Cancer Epidemiology Biomarkers & Prevention*, vol. 17, no. 11, 2008.

[7] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," *ACM Transactions on Computer Systems (TOCS)*, vol. 26, no. 2, 2008.

[8] The Apache Software Foundation, "Apache Hadoop." `hadoop.apache.org`.

[9] M. Zaharia, T. Das, H. Li, T. Hunter, S. Shenker, and I. Stoica, "Discretized streams: Fault-tolerant streaming computation at scale," in *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, ACM, 2013.

[10] The Apache Software Foundation, "Apache Mahout." `mahout.apache.org`.

[11] M. Streit, *Metabolic Pathways Influencing Gene-Expression Analysis*. Master's thesis, Graz University of Technology, 2007.

[12] R. Miller, "Response time in man-computer conversational transactions," *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, 1968.

[13] L. A. Bongo, E. Pedersen, and M. Ernstsen, "Data-intensive computing infrastructure systems for unmodified biological data analysis pipelines," *Proceedings of CIBB*, 2014.

[14] Kanehisa Laboratories, "KEGG: Kyoto Encyclopedia of Genes and Genomes." `kegg.jp`.

[15] D. Nishimura, "BioCarta," *Biotech Software & Internet Report*, vol. 2, 2001.

[16] Z. Hu, J. Mellor, J. Wu, and C. DeLisi, "VisANT: an online visualization and analysis tool for biological interaction data.," *BMC bioinformatics*, vol. 5, 2004.

[17] B. H. Junker, C. Klukas, and F. Schreiber, "VANTED: a system for advanced data analysis and visualization in the context of biological networks.," *BMC bioinformatics*, vol. 7, 2006.

[18] J. M. Villaveces, R. C. Jimenez, and B. H. Habermann, "KEGGViewer, a BioJS component to visualize KEGG Pathways," *F1000Research*, vol. 3, 2014.

[19] C. Partl, D. Kalkofen, A. Lex, K. Kashofer, M. Streit, and D. Schmalstieg, "enRoute: Dynamic path extraction from biological pathway maps for in-depth experimental data analysis," in *2012 IEEE Symposium on Biological Data Visualization (BioVis)*, IEEE, Oct. 2012.

[20] A. Lex, C. Partl, D. Kalkofen, M. Streit, S. Gratzl, A. M. Wassermann, D. Schmalstieg, and H. Pfister, "Entourage: visualizing relationships between biological pathways using contextual subsets.," *IEEE transactions on visualization and computer graphics*, vol. 19, 2013.

[21] M. Streit, M. Kalkusch, K. Kashofer, and D. Schmalstieg, "Navigation and Exploration of Interconnected Pathways," *Computer Graphics Forum (EuroVis '08)*, vol. 27, 2008.

[22] N. Kono, K. Arakawa, R. Ogawa, N. Kido, K. Oshita, K. Ikegami, S. Tamaki, and M. Tomita, "Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API.," *PloS one*, vol. 4, 2009.