SOFTWARE TOOL ARTICLE

REVISED Kvik: three-tier data exploration tools for flexible analysis of genomic data in epidemiological studies [version 2; referees: 1 approved, 2 approved with reservations]

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

[1]Department of Computer Science, UiT - The Arctic University of Norway, Tromsø, 9037, Norway
[2]Department of Community Medicine, UiT - The Arctic University of Norway, Tromsø, 9037, Norway

v2 **First published:** 30 Mar 2015, **4**:81 (doi: 10.12688/f1000research.6238.1)
Latest published: 16 Jun 2015, **4**:81 (doi: 10.12688/f1000research.6238.2)

Abstract

Kvik is an open-source framework that we developed for explorative analysis of functional genomics data from large epidemiological studies. Creating such studies requires a significant amount of time and resources. It is therefore usual to reuse the data from one study for several research projects. Often each project requires implementing new analysis code, integration with specific knowledge bases, and specific visualizations. Although existing data exploration tools are available for single study data exploration, no tool provides all the required functionality for multistudy data exploration. We have therefore used the Kvik framework to develop Kvik Pathways, an application for exploring gene expression data in the context of biological pathways. We have used Kvik Pathways to explore data from both a cross-sectional study design and a case-control study within the Norwegian Women and Cancer (NOWAC) cohort. Kvik Pathways follows the three-tier architecture in web applications using a powerful back-end for statistical analyses and retrieval of metadata.In this note, we describe how we used the Kvik framework to develop the Kvik Pathways application. Kvik Pathways was used by our team of epidemiologists toexplore gene expression data from healthy women with high and low plasma ratios of essential fatty acids.

Open Peer Review

Referee Status: ✓ ? ?

	Invited Referees 1	2	3
REVISED version 2 published 16 Jun 2015	✓ report		? report
version 1 published 30 Mar 2015	? report	? report	

1 **Paul Klemm**, Otto-von-Guericke University Magdeburg Germany

2 **Zhenjun Hu**, Boston University USA

3 **Lilit Nersisyan**, National Academy of Sciences of Armenia Armenia

Discuss this article

Comments (0)

Corresponding author: Lars Ailo Bongo (larsab@cs.uit.no)

How to cite this article: Fjukstad B, Standahl Olsen K, Jareid M *et al.* **Kvik: three-tier data exploration tools for flexible analysis of genomic data in epidemiological studies [version 2; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2015, **4**:81 (doi: 10.12688/f1000research.6238.2)

Grant information: This work was supported by a grant from the European Research Council, under the title "Transcriptomics in cancer epidemiology - TICE".

Competing interests: No competing interests were disclosed.

First published: 30 Mar 2015, **4**:81 (doi: 10.12688/f1000research.6238.1)

REVISED **Amendments from Version 1**

Overall we reduced the implementation details in the note. This was something that both reviewers pointed out and we felt that the note was a bit too technical. We also clarified the difference between Kvik and Kvik Pathways. We have changed the requirements and included a list of contributions. We also revisited the figures to make them more clear to the reader. We also fixed some grammatical errors.

See referee reports

Introduction

Visual explorative analysis is essential for understanding biological functions in large-scale omics' datasets. However, enabling the inclusion of omics' data in large epidemiological studies requires collecting samples from thousands of people at different biological levels over a long period of time. It is therefore usual to reuse the data for different research questions and projects. Although an existing tool may be useful for one project, no tool provides the required functionality for several different projects.

We have designed and implemented Kvik, a framework that makes it easy to develop new applications to explore different research questions and data. The initial version Kvik[1] contained a prototype system for exploring biological pathways and gene expression data. From this prototype we built the Kvik Framework, which provides developers a simple interface to powerful systems for statistical analyses and meta-databases, and Kvik Pathways: a publicly available data exploration application. From our experience in developing a framework for building data exploration applications, we identified four requirements such applications should satisfy:

Interactive The applications should provide interactive exploration of datasets through visualizations and integration with relevant information. To understand the large quantities of heterogeneous data in epidemiological studies, researchers need interactive visualizations that provide different views and presentations of the data. Also, to understand the results it is important to have instant access to existing knowledge from online databases.

Familiar They should use familiar visual representations to present information to researchers. For more efficient data exploration it is effective to use representations that researchers are familiar with both from the literature and from other applications.

Simple to use Researchers should not need to install software to explore their data through the applications. The applications should protect the researcher from the burden of installing and keeping an application up to date.

Lightweight Data presentation and computation should be separated to make it possible for researchers to explore data without having to have the computational power to run the analyses. With the growing rate data is produced at, we cannot expect that researchers have the resources to store and analyze data on their own computers.

There are several tools for exploring biological data in the context of pathways, such as VisANT (available online at visant.bu.edu) by 2, VANTED (available online at vanted.ipk-gatersleben.de)[3], enRoute by 4 or Entourage by 5 (both available online at caleydo.org). However, these tools do not provide the adaptability needed for exploration of multi-study datasets. Many existing tools place the visualization, data analysis and storage on the user's computer, making it necessary to have a powerful computer. In addition, the tools are often standalone applications that require users to install and update the applications. Kvik Pathways satisfies the above requirements as follows:

Interactive Kvik Pathways provides interactive pathway visualizations and information from the popular Kyoto encyclopedia of genes and genomes (KEGG)[6] database (available online at kegg.jp).

Simple to use Kvik Pathways uses HTML5 and modern JavaScript libraries to provide an interactive application that runs in any modern web browser.

Familiar Kvik Pathways uses the familiar pathway representations from KEGG and graphical user interfaces found in modern web applications.

Lightweight Kvik Pathways uses a powerful back-end provided by the Kvik framework to perform statistical analyses.

Both Kvik and Kvik Pathways are open-sourced at github.com/fjukstad/kvik. We provide an online version of Kvik Pathways at kvik.cs.uit.no and to run Kvik Pathways in a local Docker instance or on a cloud service such as Amazon Web Services (aws.amazon.com) or Google Compute Engine (cloud.google.com/compute), we provide a Docker image at registry.hub.docker.com/u/fjukstad/kvik.

In this note we describe how we used Kvik to implement Kvik Pathways, a tool for exploring gene expression in the context of biological pathways. In Kvik Pathways researchers can explore gene expression data from 7 combined with information from online knowledge bases. We provide the following contributions:

- Kvik Pathways, a publicly available web application for exploring gene expression data in the context of biological pathways without any additional applications than a web browser.
- A requirement analysis for interactive exploration tools for epidemiological studies.
- A detailed description of how we have used Kvik Pathways to explore gene expression data from healthy women with high and low plasma ratios of essential fatty acids.

Methods

Kvik Pathways allows users to interactively explore a molecular dataset, such as gene expression, through a web application. It provides pathway visualizations and detailed information about genes and pathways from the KEGG databases (Figure 1). Through pathway visualizations and integration with the KEGG databases, epidemiologists can perform targeted exploration of pathways and genes

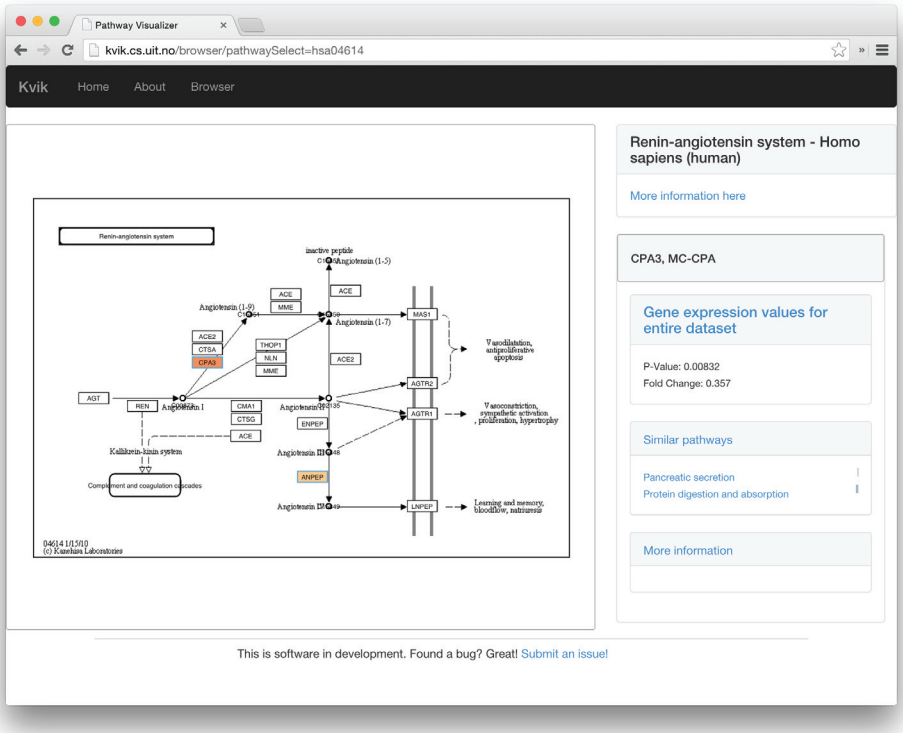

Figure 1. Screenshot of the renin-angiotensin pathway (KEGG pathway id hsa04614) in Kvik Pathways. The user has selected the gene CPA3, which brings up the panel on the right. From here researchers can browse pathways that the gene is a member of, and read relevant information about the gene from KEGG.

to get an overview of the biological functions that are involved with gene expression from the underlying dataset. Kvik Pathways gathers information about related pathways and retrieves relevant information about genes, making it unnecessary for researchers to spend valuable time looking up this information manually. For example, navigating a set of pathways and browsing information about genes in these, requires the researcher to manually query KEGG for each specific gene. Kvik Pathways retrieves information about genes without the researcher having to leave the pathway visualization to retrieve relevant information.

The Kvik framework provides a flexible statistics back-end where researchers can specify the analyses they want to run to generate data for later visualization. For example, in Kvik Pathways we retrieve fold change for single genes every time a pathway is viewed in the application. These analyses are run ad hoc on the back-end servers and generates output that is displayed in the pathways in the client's web browser. The data analyses are implemented in a simple R script and can make use of all available libraries in R, such as Bioconductor (bioconductor.org).

Researchers modify this R script to, for example, select a normalization method, or to tune the false discovery rate (FDR) used to adjust the *p*-values that Kvik Pathways uses to highlight significantly differentially expressed genes. Since Kvik Pathways is implemented as a web application and the analyses are run ad hoc, when the analyses change, researchers get an updated application by simply refreshing the Kvik Pathways webpage.

Implementation

We implemented interactive visualizations using the Cytoscape.js (js.cytoscape.org) library to generate the interactive pathway visualizations, and D3 (d3js.org) for Document Object Model (DOM) manipulation such as generating bar charts with HTML <*svg*> elements. We integrate these with the popular Bootstrap front-end framework (getbootstrap.com) to provide a familiar and aesthetically pleasing user interface.

Kvik Pathways has a three-tiered architecture of independent layers (Figure 2). The browser layer consists of the web application for exploring gene expression data and biological pathways. A front-end layer provides static content such as HTML pages and stylesheets, as well as an interface to the data sources with dynamic content such as gene expression data or pathway maps to the web application. The back-end layer contains information about pathways and genes, as well as computational and storage resources to process genomic data such as the NOWAC data repository. The Kvik framework provides the components in the back-end layer.

In our setup the Data Engine in the back-end layer provides an interface to the NOWAC data repository stored on a secure server on our local supercomputer. In Kvik Pathways all gene expression data is

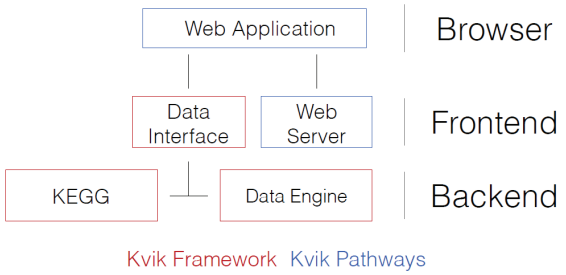

Figure 2. The three-tiered architecture of Kvik Pathways.

stored on the computer that runs the Data Engine. The Data Engine runs an R session accessible over remote procedure calls (RPCs) from the front-end layer using RPy2 (rpy.sourceforge.net) to interface with R. To access data and run analyses the Data Interface exposes a HTTP API to the browser layer (Table 1 provides the interfaces).

To create pathway visualizations the Kvik back-end retrieves and parses the KEGG Markup Language (KGML) representation and pathway image from KEGG databases through its REST API (rest.kegg.jp). This KGML representation of a pathway is an XML file that contains a list of nodes (genes, proteins or compounds) and edges (reactions or relations). Kvik parses this file and generates a JSON representation that Kvik Pathway uses to create pathway visualizations. Kvik Pathways Cytoscape.js to create a pathway visualization from the list of nodes and edges and overlay the nodes on the pathway image. To reduce latency when using the KEGG REST API, we cache every response on our servers. We use the average fold change between the groups (women with high or low plasma ratios of essential fatty acids) in the dataset to color the genes within the pathway maps. To highlight *p*-values, the pathway visualization shows an additional colored frame around genes. We visualize fold change values for individual samples as a bar chart in a side panel. This bar chart gives researchers a global view of the fold change in the entire dataset.

Operation

Kvik Pathways runs in all modern web browsers and does not require any third-party software.

Table 1. The REST interface to the Data Engine. All URLs are relative to the hostname where the Data Engine server runs. On our public installation the Data Engine runs on kvik.cs.uit.no:8888. For example, use kvik.cs.uit.no:8888/genes/ to retrieve all available genes in our dataset. By using a HTTP API we can build different data exploration applications in virtually any programming language.

URL	Description
/fc/[genes...]	Calculate and retrieve fold-change for the specified genes
/pvalues/[genes...]	Calculate and retrieve *p*-values for the specified genes
/exprs/[genes...]	Get the raw gene expression values from the dataset
/genes	Get a list of all genes in the dataset

Use case

We used Kvik Pathways to repeat the analyses in a previous published project (7, doi: 10.1371/journal.pone.0067270) that compared gene expression in blood from healthy women with high and low plasma ratios of essential fatty acids. Gene expression differences between groups were assessed using *t*-tests (*p*-values adjusted with the Benjamini-Hochberg method). There were 184 differentially expressed genes significant on the 5% level. When exploring this gene list originally, functional information was retrieved from GeneCards and other repositories, and the list was analyzed for overlap with known pathways using MSigDB (available online at broadinstitute.org/gsea/msigdb). The researchers had to manually maintain overview of single genes, gene networks or pathways, and gather functional information gene by gene while assessing differences in gene expression levels. With this approach, researchers are limited by manual capacity, and the results may be prone to researcher bias. Kvik Pathways eliminates this researcher bias and does not limit the information retrieval to a researcher's manual capacity.

Initially, Kvik Pathways was implemented to explore gene expression data from a not yet published dataset. To use Kvik Pathways to explore the data from the analyses in 7, we only needed to make small modifications to the analysis R script used by the Data Engine. (The modified R script is found at github.com/fjukstad/kvik/blob/master/dataengine/data-engine.r). Instead of loading the unpublished dataset, we could load the dataset from 7 and use the four functions that are accessible over RPC (Table 1 shows the HTTP API which uses the underlying RPCs). Currently this script is less than 30 lines, consisting of four functions to retrieve data and a simple initialization step that reads the dataset. Researchers only have to modify these four functions to enable exploration of new datasets. As of the current implementation of Kvik Pathways researchers have to modify the analysis script outside the application.

As an example of practical use of Kvik Pathways, we chose one of the significant pathways from the overlap analysis, the renin-angiotensin pathway (Supplementary table S5 in 7). The pathway contains 17 genes, and in the pathway map we could instantly identify the two genes that drive this result. The color of the gene nodes in the pathway map indicates the fold change, and the statistical significance level is indicated by the color of the node's frame. We use this image of a biological process to see how these two genes (and their expression levels) are related to other genes in that pathway, giving a biologically more meaningful context as compared to merely seeing the two genes on a list.

Summary

Kvik Pathways is an open-source system for explorative analyses of functional genomics data from epidemiological studies. It uses R to perform on-demand data analyses providing a flexible back-end that can expand to new analyses and research projects. It uses modern visualization libraries and a powerful back-end for on-demand statistical analyses. Epidemiologists are using Kvik Pathways to analyze gene expression data. Kvik Pathways is open-sourced at github.com/fjukstad/kvik and is available as a Docker image at registry.hub.docker.com/u/fjukstad/kvik.

Data availability

Data used in the use case is available in the Gene Expression Omnibus (ncbi.nlm.nih.gov/geo), under accession number GSE15289.

Software availability

Latest source code

https://github.com/fjukstad/kvik

Source code as at the time of publication

https://github.com/F1000Research/kvik/releases/tag/1.0

Archived source code as at the time of publication

http://dx.doi.org/10.5281/zenodo.16375

Software license

The MIT license.

Author contributions

LAB and BF designed the architecture of the system. BF implemented. All conducted the requirements analysis. EL, MJ, KSO contributed case study. BF drafted manuscript. All authors read, revised and approved the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by a grant from the European Research Council, under the title "Transcriptomics in cancer epidemiology - TICE".

Acknowledgements

Gene expression profiles were analyzed at the Microarray Resource Center Tromsø, UiT – The Arctic university of Norway.

References

1. Fjukstad B, Olsen KS, Jareid M, *et al.*: **Kvik: Interactive exploration of genomic data from the NOWAC postgenome biobank.** Norsk Informatikkonferanse (NIK). 2014.
Reference Source
2. Hu Z, Chang YC, Wang Y, *et al.*: **VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies.** *Nucleic Acids Res.* 2013; **41**(Web Server issue): W225–W231.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Junker BH, Klukas C, Schreiber F: **VANTED: a system for advanced data analysis and visualization in the context of biological networks.** *BMC Bioinformatics.* 2006; **7**(1): 109.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Partl C, Lex A, Streit M, *et al.*: **enRoute: Dynamic path extraction from biological pathway maps for in-depth experimental data analysis**. In *Biological Data Visualization (BioVis), 2012 IEEE Symposium on,* pages 107–114.
Publisher Full Text
5. Lex A, Partl C, Kalkofen D, *et al.*: **Entourage: visualizing relationships between biological pathways using contextual subsets.** *IEEE Trans Vis Comput Graph.* 2013; **19**(12): 2536–2545.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res.* 2000; **28**(1): 27–30.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Olsen KS, Fenton C, Frøyland L, *et al.*: **Plasma fatty acid ratios affect blood gene expression profiles--a cross-sectional study of the Norwegian Women and Cancer Post-Genome Cohort.** *PLoS One.* 2013; **8**(6): e67270.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status:

Version 2

Referee Report 26 August 2015

doi:10.5256/f1000research.7113.r10131

Lilit Nersisyan
Group of Bioinformatics, Institute of Molecular Biology, National Academy of Sciences of Armenia, Yerevan, Armenia

In their paper, Fjukstad *et al.,* have described Kvik, a web-based tool intended for KEGG pathways based analysis of gene expression data, and easy sharing and delivery of analysis results to a wide audience via web-based visualization.

General feedback:

The authors have tried to separate the knowledge base, from heavy statistical calculations, and those from actual visualization of the results. This would make the software both flexible for modifications, adjustable for growing data, and at the same time, could allow the end user to see the updated version of calculations by simply updating their browser.

While this idea is of great value, the implementation seems not to fully correspond to the intent, and the manuscript fails to deliver the proper functionality, the applications, and the target audience of the software.

I had to refer to the authors' previous publication, "Kvik: Interactive exploration of genomic data from the NOWAC postgenome biobank" (conference paper), to understand the actual intent of the framework and the Kvik pathways. To my understanding, the current paper adds little value to this previous publication, even with the use case described.

The authors could stress on the actual functionality of Kvik, describe more elaborate use cases, and compare it to existing approaches. The advantages of the tool are not clear: those are mentioned in the introduction and the summary, but do not really correspond to the actual implementation state, and are not addressed in the text body.

***Title**:*

The title refers to the *general* idea of three-tier data exploration tools and their value for *general* use in epidemiological studies. However, the paper describes only Kvik pathways developed with their framework, and only a simple mapping of gene expression data onto KEGG pathways.

Abstract:

The abstract concentrates on the problem of multiple usages of the same data in different studies, and mentions that Kvik somehow solves this problem. There is no at least understandable message in the text describing how exactly Kvik is a solution.

Article content:

The content was hard to comprehend. I'd like to mention the following major points that might be relevant:

1. The introduction section is redundant with double notice of the features a good framework should satisfy and which features Kvik itself possesses. Probably the authors could concentrate more on what features they intended to incorporate in Kvik, and why those features are important, in the same paragraph, and describe how exactly these features are satisfied further in the implementation.

2. The introduction itself messes up the target audience: in the beginning it mentions that Kvik is to provide a simple framework for developers to apply advanced statistics on the data, but later on mentions only that the 'easy to use' feature of Kvik is for the end-users to see the visualization results.

3. As continuation of the previous comment, the paper fails to describe how exactly a user can modify the statistical analysis algorithms in the R scripts ("outside the application" is not sufficient). This seems to be one of the main features of Kvik, and it is mentioned in the introduction and mostly in the summary, however no further explanations or examples are provided in the body.

4. It is not true that having a browser is everything that is needed for using Kvik. It's true for exploration of Kvik pathways only, but not for applying custom statistical analysis. Additionally, the users need to be familiar with how to use Docker images, and this again poses the question of who is the target audience.

5. It is confusing to load pathways at http://kvik.cs.uit.no/, and see FC values without knowing what these FC values actually mean. The user does not need to refer to the paper to see which values of which dataset are compared with each other to derive those FC values, and the browser only gives the information that the data is from the NOWAC biobank. Again, the paper fails to fully describe how the R script should be modified to include their own dataset and how these modifications will appear on the browser.

6. It is too slow. With 13 Mbps internet speed, it requires more than 30 seconds to load the list of pathways, and nearly 30 seconds to load a single pathway.

7. In the use case, the comparison of Kvik with the previously published approach is not appropriate. The authors do not clearly distinguish between gene set enrichment and pathway overrepresentation (overlap) analysis. The fact that the researchers should manually lookup gene functions is not solved by Kvik, since the output of Kvik is not the same as the output of those researchers: it simply provides mapping of expression data onto pathways. Moreover, the notion of "restricted to researcher's manual capacity and the results may be prone to researcher bias" is not addressed by Kvik. Even though the genes appear in the pathway automatically, the researcher still has to scroll through the pathways and subjectively decide what the output means in biological sense. Thus, the output is limited and the bias is still there.

8. There is no comparison with existing tools, except for the sentence *"However, these tools do not provide the adaptability needed for exploration of multi-study datasets"* in the introduction, which is not clear.

Summary

The summary is a good overview of the authors' initial intent and idea, which is great. However, the summary does not really correspond to the body of the text.

Minor points

1. The browser is unresponsive when in the middle of loading one wants to push the back button.

2. Too little data. The majority of genes in the pathways do not have associated FC values. This makes the visualization results not clear. The authors may either mention this in the paper, or provide a richer dataset.

3. It is not clear what is meant by:

    1. "*required functionality for multistudy data exploration*" in the abstract.

    2. "*from thousands of people at different biological levels*" in the introduction.

    3. *"A requirement analysis for interactive exploration tools*" in the introduction.

4. The authors should make clear what fold change values are used exactly, when they mention them in the text. E.g. in the "*For example, in Kvik Pathways we retrieve fold change for single genes*" part in the second paragraph of the Methods section.

5. When mentioning bar charts in the Implementation section, the authors can provide a figure with a bar chart, or provide a respective link.

6. In the Use case, it is mentioned that *"the statistical significance level is indicated by the color of the node's frame"*. However, the color code is not clear and it is also not mentioned in the Figure 1 legend, nor there is any information found in the Kvik pathways.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Referee Report 25 June 2015

doi:10.5256/f1000research.7113.r9080

Paul Klemm
Department Simulation and Graphics, Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

All criticized points were considered in the revision. I think the paper has substantially improved. The distinction between `Kvik` and `Kvik Pathways` is made clear and a list of contributions makes it easier to follow the structure. I think you should consider putting the requirement analysis into a dedicated section.

I see no reason to not approve the paper in its current state.

I have noticed two minor typos:

- Abstract
  - Typo: Missing whitespace after sentence "Kvik Pathways follows the three-tier architecture in web applications using a powerful back-end for statistical analyses and retrieval of metadata."
- Implementation
  - Missing verb? "Kvik Pathways Cytoscape.js to create a pathway visualization from the list of nodes and edges and overlay the nodes on the pathway image."

Apart from the paper, I have one comment on your response on my review w.r.t. the differences regarding the NIK paper. You stated that
"*we removed the security since we believe that data should be publicly available, [...]*"

I completely agree with the notion of open data. In my experience though, epidemiologists are often bound to contracts enforcing confidentiality of the data to protect the privacy of the participants. Our project partners legally must not use a system, which does not allow for secure data handling.

This has no consequence for your paper at all, I just think that this is a observation from our collaboration with epidemiologists you might appreciate.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 30 April 2015

doi:10.5256/f1000research.6693.r8533

Zhenjun Hu
Bioinformatics Graduate Program and Department of Biomedical Engineering, Boston University, Boston, MA, USA

The manuscript presents Kvik as an open-source system for explorative analysis of functional genomics data from large epidemiological studies. The authors seem have excellent ideas, but the implementation of the tool is far behind these ideas. I would like to approve the manuscript if the following points can be addressed:

1. The target of the tools. There are in general two type of tools: data provider and tool provider. The two of course can be combined. The prior in general provides knowledge, and the later provides functions to analyze users' own data. Kvik however seems lack the data to be a knowledge

provider, and also does not provide enough functionality to be the later. To be the former, I will recommend authors to add more epidemiological data, to be the later, the author need to give clear instruction how user's own data can be analyzed using Kvik. For an example, the idea to connect to the cloud service is excellent, but how can Kvik to achieve this?

2. The implementation of Kvik seems to be improved, especially the performance. When I tried the Kvik, the browser tells me several time that the page is not responding. Yet I know the page is responding, but just take too much time. In addition, from data security point of view, it is not good to use RPCs from the browser layer to data engine directly, it shall be avoid in general in the three-tier architecture.

3. The manuscript need to focus more on the functionality of the tool. The current manuscript has too many technical details.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 02 Jun 2015

Bjørn Fjukstad, University of Tromso, Norway

We would first like to thank the reviewer Zhenjun Hu for his thorough feedback and comments.

1. Since we have open-sourced the application we believe that Kvik Pathways can provide knowledge in itself, but also be used by other researchers to gain knowledge in their data. We have modified the use case to make it more clear how users can do so themselves. We also refer to the online repository where this information is also available. As we state in the Introduction we provide Docker images for running Kvik Pathways in a cloud service.

2. We agree, the current implementation can be improved. We are working on a second version where we have reduced the latency and that gives more feedback to the user if he has to wait. Regarding using RPCs from the browser layer we agree. We have updated the note with more details on how the system is implemented. The Data Engine provides an HTTP REST API that the browser layer queries. It does not send RPC requests from the browser layer, but from the frontend when it receives a request to the HTTP REST API. Also, the only RPCs that are allowed to run are defined in the R analysis script.

3. We agree. We have reduced the number of implementation details to make the functionality stand out more.

Competing Interests: No competing interests were disclosed.

Referee Report 23 April 2015

doi:10.5256/f1000research.6693.r8165

Paul Klemm
Department Simulation and Graphics, Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

The work presented by Fjukstad *et al*. pursues in pushing the notion of open science in epidemiology. It describes Kvik, a web-based tool for analyzing genomic pathways. I really like the ideas behind it and value the detailed implementation section as well as the state-of-the-art techniques used. On the other hand I think that the paper should focus more on describing the epidemiological context, associated requirements and target groups to communicate the design choices for Kvik. Go into detail on the application *workflow.*

General Feedback

I like the approach the authors take with the paper. The tool they describe seems to be well suited for analyzing genomic pathways in the epidemiological context.

I miss a clear statement on the papers contribution. Maybe you can put a bullet list in the *Introduction* section to tell the reader what things can be done with your software, which could not be done before!

In my understanding, epidemiology is a very interdisciplinary in terms of associated experts. You have your clinicians deriving hypotheses from their day to day practice, statisticians deriving statistically sound conclusion as well as biologists and computer scientists associated with such projects. *Which of these are your target group?*

When you described your target users, describe what they are trying to find out. How does your tool help them doing that? Does it allow them do their work faster? Do they derive insights they could not get before? The latter would be a huge contribution! Please give more details on the *workflow* of you system!

I miss a clear distinction from the NIK-2014 paper "Kvik: Interactive exploration of genomic data from the NOWAC postgenome biobank". The paper was also not cited in this work. It seems to me that the majority of the content of the presented paper can already be found in the NIK-2014 paper. Please elaborate on the differences and cite the paper. *If you can not state clear differences, there is, in my opinion, no point in publishing this paper and I will rate it 'Not Approved'*.

Title
The title of the work is appropriate.

Abstract
The abstract motivates the need for new tools, which allow to assess the vast amount of epidemiological data well. In my opinion it can be improved by:

- reduce the amount of implementation detail. You tell the reader later on which frameworks and libraries you use
- explain who are your users.
- what can be done with your tool now, which could not be done before?

Minor comments on the abstract:

- "Existing data exploration tools do not provide all the required functionality for such multi-study data exploration." This is a dangerous statement, since you do not say anything about what the

required functionality is! I think I know what you are trying to get at, in the introduction you describe it better with: "Although an existing tool may be useful for one project, no tool provides the required functionality for several different projects."

Introduction
The introduction can be improved by clearly stating the contributions (e.g., as bullet points).

I would like to see some reference or a method on how the five requirements were acquired. These are all things, which are important in almost all applications. Where is the difference of software in the epidemiological context towards other context and how does Kvik adapt to the arising requirements? You answer many of these questions, later on when you repeat all the requirements again, but to me it is not structured well.

Methods
The method section is written well. I would like to know how the users modify the R scripts (beginning second paragraph). Do they do this inside Kvik or do they have to switch into another software for it?

Figure 1 caption: What can the user do now after he or she selected the gene? The workflow is not clear to me.
Figure 2 was already presented very similarly in the NIK'14 paper.

Minor: Closing parenthesis in sentence "In our setup the Data Engine in the back-end layer provides an interface to the NOWAC data repository stored on a secure server on our local Stallo Supercomputer Table 1 provides the interfaces)."

Use Case
The use case section can be strengthened by reducing the amount of implementation details (in my opinion mentioning the individual function names is not necessary to comprehend the functionality) and focusing more on the involved actors and tasks and contexts associated with the use case. What feedback was given by the user(s)?

Reusability
The effort of the authors to make the software publicly available is worth a special note. Modern state of the art techniques are combined with powerful back-end systems, which scale well on different application scenarios.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 02 Jun 2015

Bjørn Fjukstad, University of Tromso, Norway

We would first like to thank the reviewer Paul Klemm for his thorough feedback and comments.

Difference between the NIK paper and the Application Note

When we wrote the NIK paper, Kvik was only a system for exploring genomic data in the context of

pathways. Since then we have realized that exploring genomic data in pathways is not enough, that we need a framework that allows us to build different applications for exploring data from different studies with different designs. With this in mind we have refined the requirement analysis from our initial Kvik system and developed more general requirements that these applications should satisfy.

From our initial Kvik implementation we have now decoupled the application (Kvik Pathways) from the framework, allowing fast development of new applications. The Kvik framework provides interfaces to a Data Engine that provides statistical analyses, and interfaces to online databases such as KEGG. Kvik Pathways is the first application that we developed using the Kvik Framework. Using Kvik we have developed several other applications that will be published in the near future.

In the NIK paper we described from a computer science point of view the features of Kvik, both looking at the application itself, and the backend features that are now a apart of the Kvik Framework. The NIK paper was written to give a more in detail view of how the system works and performs, while in this application note we want to describe how our epidemiology researchers helped to develop the application and how they used it to reproduce results they found in an already published dataset.

Using an already published dataset was important to us since it allows us to provide a publicly available Kvik installation for others to use. We will revisit the second paragraph of the Use case section where we discussed how we used the initial Kvik system to explore different data from a different study design.
To sum up, our new contributions in the application note are as follows:

- Publicly available application
- Publicly available Docker containers that researchers can use to set up local installations of Kvik Pathways.
- Reproduced the results from an already published dataset to make the system publicly accessible at kvik.cs.uit.no
- A more refined requirement analysis that reflects our experiences after publishing the NIK paper. The important changes are:

  i) emphasis on integration of online knowledge bases (interactive requirement),

  ii) emphasis on the system being flexible to adapt to data and different statistical analyses,

  iii) we removed the security since we believe that data should be publicly available, and

  iv) put emphasis on separating computation and visualization (lightweight).

We have cited the NIK paper in the application note and improved the text to highlight the differences between the framework and the Kvik Pathways application.

We have included a list of contributions in the Introduction section.

We agree that epidemiology is a very interdisciplinary. Kvik Pathways has been developed in a team of epidemiologists, biologists, statisticians and computer scientists. The application note

targets such groups of researchers working together to develop systems for gaining biological insights in genomic data. We have re-written parts of the note to clarify how researchers have used the application.

Abstract : We agree and have reduced the amount of implementation detail and made it more specific what our users have done with the application.
Introduction: We have revisited the requirements and specified how these are different from regular applications. We believe that it is best to separate the requirements and how we solved them in two different lists.

Methods: As of today users modify R scripts outside Kvik. We have made it clear in the methods section. We will expand Figure 1 caption to clearify the workflow. Regarding figure 2. We chose to include it since it highlights the important three-tiered architecture with applications that use the Kvik framework. We have modified the figure to highlight connection between the application and the framework.

Use case: As mentioned we will expand this section with a more detailed workflow. Since it is an app note targeted towards users we will reduce implementation details and refer to the source code.

Competing Interests: No competing interests were disclosed.