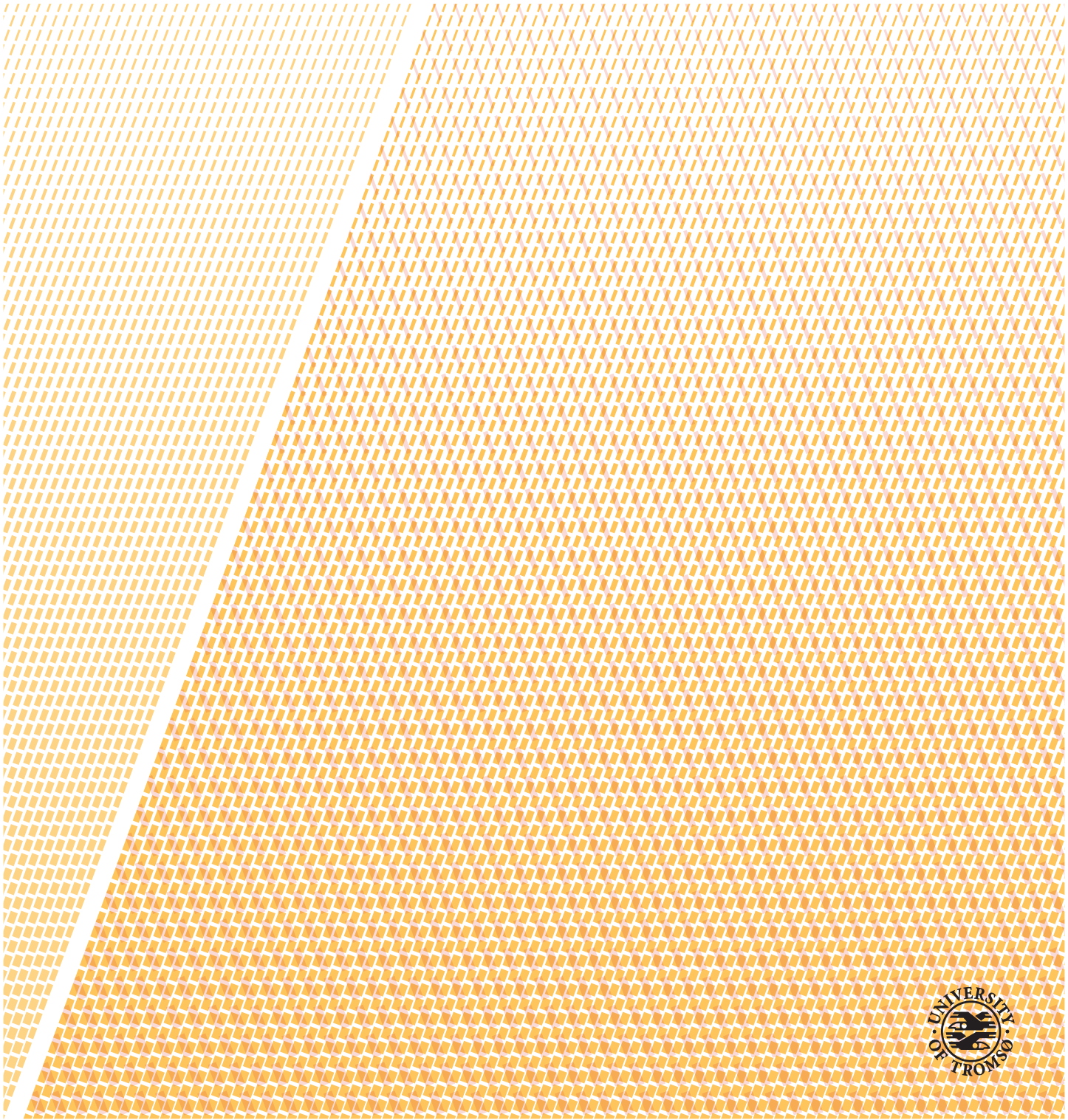
UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Faculty of Science and Technology
Department of Computer Science

A Unified Approach for Reproducible Analysis of High-Throughput Biological Datasets

Bjørn Fjukstad

A dissertation for the degree of Philosophiae Doctor



Dear everyone, I'm sorry.

Abstract

Insert abstract here.

Acknowledgements

Insert acks here.

Contents

1

Introduction

There is a rapid growth in the number of available biological datasets due to the decreaseing cost of data collection. This brings opportunities to gain novel insights to the underlying biological mechanisms in the development and progression of diseases such as cancer, possibly leading to the development of novel diagnostic tests and drugs for treatment. The wide range of different biological datasets has led to the development of a wealth of software packages and systems to explore and analyze these datasets. However, there are few tools toat are designed with the full analysis pipeline in mind, from raw data into interpretable results. While the tools are used to provide novel insights in diseases, there is little emphasis on reporting and sharing information about tool versions, input parameters, and other information that can help others use the same known methods on their own datasets. This leads to unneccesary difficilties to reuse known methods, and difficulties in reproducing analyses, which leads to longer analysis times and unrealized potential for scientific insights.

There are several computational challenges challenges for researchers to analyze and explore biological datasets. These challenges are common for large datasets such as high-throughput sequencing data that require long-running, deep analysis pipelines, as well as smaller datasets, such as microarray data, that require complex, but short-running analysis pipelines. The first is the time and knowledge required to find and set up the necessary analysis tools to start analyzing a modern biological dataset. The second is ensuring the correct input parameters, tool versions, database versions, and dataset versions

when analyzing, and reporting analysis results to enable reproducible science. A third challenge is efficiently exploring the results of the analyses interactively. This includes developing tools that can efficiently visualize the heterogeneous datsets and integrate them with known biology from databases to provide necessary information for interpreting the results. The final challenge is reusing the analysis pipelines and exploration tools with new datasets, methods, and research questions.

As a result, there are a wealth of specialized approaches and systems to enable analysis of the complex biological data. To develop deep analysis pipelines in bioinformatics, Galaxy[1] has for a long time provided a simple interface to set up and execute analysis pipelines for genomic datasets. However, the Galaxy system is less effective for explorative and flexible analyses where it is necessary to try out different tools with different configurations.[2] New initiatives such as the Common Workflow Language (CWL) provide users a standardized way of describing and sharing an analysis pipeline, and has multiple implementations such as the reference implementation cwl_runner,[1] Arvados,[3] Rabix,[4] Toil,[5] Galaxy,[1] and AWE.[6] While these systems provide a viable option for batch-processing of large biological datasets, researchers with smaller datasets at hand, can analyze and explore them through interactive languages and interpreters such as Python or the R programming language. Through the package repository Bioconductor, there are a wide range of R packages to analyze biological datasets. These include tools for both analyzing and visualizing the datasets. For users with little or no programming experience it is possible to access and explore datasets thorough applications using the Shiny or OpenCPU frameworks. These let developers write applications in the R programming language, and users can access and explore the data through web applications. Standalone systems such as Cytoscape provide a specialized software platform to visualize and explore complex biological datasets.[**?**] Generalized systems for analyze a wide range of big datasets are now starting to get attention in bioinformatics. An example of one of these is Pachyderm, a system for deploying and managing multi-stage, lagnuage-agnostic data pipelines.[**?**] In addition to tracking pipeline configrations, it also provides full provenance for the data. Another example is Aparche Spark, an analytics engine for large-scale data processing.[**?**]. Both of these provide useful abstractions for reproducible analyses of large-scale datasets, but they have yet to see wide-spread adpotion in Bioinformatics. With the addition of new datasets and methods every year, it seems that analysis of biological data requires a wide array of different tools and systems.

This dissertation argues that, instead, we can design a unified approach that integrates disperate systems and data into fully reproducible biological data anal-

1. `github.com/common-workflow-language/cwltool`

ysis frameworks. In particular, we show how software container technologies together with well-defined interfaces, configurations, and orchestration provide the necessary foundation to build reproducible analysis pipelines for biological datasets, as well as highly interactive data exploration applications.

The resulting approach as several key advantages when implementing systems to analyze and explore biological data:

- It enables reproducible research by packaging applications and tools within containerized environments. This enables sharing of tools and simplifies the tedious task of installing specific tools.

- It simplifies the sharing of analysis pielines and workflows across different research teams and systems. This shortens the time-to-interpretation for biological datasets.

- It enables applications to use tools written in any programming language, using open standards to communicate between tools and systems. This allows for exploration tools to interface with both statistical analyses and biological databases.

- It facilitates the development of flexible and configurable systems by separating appliications and tools into small composable parts. This allows developers to reuse parts of a system to fit new methods and datasets.

From collaboration with researchers in systems epidemiology and precision medicine we were asked to develop a set of applications and systems that could enable them to analyze and explore their datasets. From these systems we extraplolated a set of general design principles to form a unified approach. We implement our approach thorugh a series of applications and tools built on top of a stack of open source systems with software containers as the common foundation. We evaluate the approach through these systems using real datasets and show its viability.

From a longer-term perspective we discuss the general patterns for implementing modern data analysis systems for use in precision medicine and dicuss why our approach is a suitable option. As more datasets are produced every year, research will depend on systems being easy to pick up, and provide the necessary functionality to reproduce and share the analysis pipelines.

Thesis statement: A unified development model based on software container infrastructure can efficiently provide reproducible and easy to use environments to develop applications for exploring and analyzing biological datasets.

1.1 Problems with Data Analysis and Exploration in Bioinformatics

Today there is a move towards using more sophisticated approaches to analyze biological datasets through workflow and pipeline mangers such as Galaxy[1] and the CWL[**?**]. These simplify setting up the analysis pipeline pipeline, maintaining, and updating it. However, these tools still have their limitations and shell scripts are still the de facto standard building analysis pipelines in bioinformatics. For exploring biological data there are a range of tools, such as Cytoscape[**?**] and Circos[**?**], that support importing an already-analyzed dataset to visualize and browse the data.

Although there are efforts to develop tools to help researchers explore and analyze biological datasets, they current tools have several drawbacks:

1. **Reusability:** Data exploration tools are often developed as a single specialized application, making it difficult to reuse parts of the application for other analyses or datasets. This leads to duplicate development effort and abandoned projects.

2. **Decoupling:** Data exploration tools are often decoupled from the statistical analyses. This often makes it a difficult exercise to document and retrace the analyses behind the results.

3. **Complexity:** Analyses that start as a simple script quickly become more difficult to maintain and develop as developers add new functionality to the analyses.

4. **Reproducibility:** While there are tools for analyzing most data types today, there is little or no effort to fully document the entire pipeline from raw data to interpretable results. This includes tool versions, parameters, data, and databases. This makes analysis results difficult to reproduce.

Because of these drawbacks, an approach for reproducible data analysis and exploration would have significant benefits for the complex interpretation of biological datasets.

1.2 The X Model/approach/etc.

From the collaboration with researchers we have developed applications for two specific research areas. The first to explore the datasets of a large population-

based research cohort. The second to analyze sequencing datasets for use in a precision medicne setting. Although these areas require widly different systems with different requirements, the systems share common design patterns. We discuss the different areas separately before highlighting the similarities.

Deep analysis pipelines. Analysis of high-throughput sequencing datasets requires deep analysis pipelines with a large number of steps that transform raw data into interpretable results[7]. There are a large number of tools available to perform the different processing steps, written in a wide range of programming languages. The tools, and their dependencies, can be difficult to install, and they require users to correctly manage a range of input parameters that affects the output results. With these observations in mind we used software containers to package the tools we needed for our analyses, one tool per container image. This made it possible to share the container image between compute systems without installing any dependencies or additional packages. To keep track of input parameters as well as the flow of data in a pipeline we designed a text-based specification for analysis pipelines. This specification includes information such as input parameters and tool versions.

This approach was then implemented in walrus, a tool that lets users create and run analysis pipelines. In addition, it tracks full provenance of the input, intermediate, and output data, as well as tool parameters. With *walrus* we have successfully built analysis pipelines to detect somatic mutations in breast cancer patients, as well as an Ribonucleic acid (RNA)-seq pipeline for comparison with gene expression datasets.

Interactive exploration. Analysis pipelines and workflows typically require researchers to browse and explore the final output. In addition it may be useful to further explore results by modifying analysis parameters to execute new analyses. As with analysis pipelines there are complete exploration tools as well as software libraries to develop custom applications. The tools often require users to import already analyzed datasets but provide interactive visualizations and point-and-click interfaces to explore the data. Users with programming knowledge can use the wealth of software packages for visualization within languages such as R or Python. With a lot modern visualization libraries created for the web there are also possibilities to develop applications that target users on any platform. From these observations we wrote an interface to the R programming language, that would allow us to interface with the wealth of existing software packages for biological data analyses from a point-and-click application. New data exploration applications could access analyses directly through this interface, removing the previous decoupling between the two. In addition, to provide reproducible execution environments we also packaged into software containers that could be easily deployed and shared.

This approach was then implemented as a part of *Kvik*, a collection of packages to develop new data exploration applications. Kvik allows applications written in any modern programming language to interface with the wealth of bioinformatics packages in the R programming language, as well as information available through online databases. We have used Kvik to develop the Matched Interactions Across Tissues (MIXT) system for exploring and comparing transcriptional profiles from blood and tumor samples in breast cancer patients, in addition to applications for exploring biological pathways.

Commonalities. Both approach to build analysis pipelines and to write interactive data exploration applications build on the same principles. In both areas we break down the systems in to smaller composable units, e.g. a tool, and package these into a software container which are then orchestrated together. These containers are configured and communicate using open protocols that make it possible to interface with them using any programming language. We can keep track of the configuration of the containers and their orchestration using software versioning systems, and provide the necessary information to reproduce analyses or a complete system.

1.3 Systems Implemented with the X Model/approach/etc.

In this section we detail the different systems we have developed using the approach.

We have used the X Model (X) to implement both batch processing systems targeted at high-throughput analysis pipelines, as well as interactive data exploration systems for interactively exploring the results and emerging patterns from these analyses. We discuss the different systems and areas we have implemented.

Combined these demonstrate how the X approach has been used for a full

1.4 Summary of Results

1.5 List of papers

This section contains a list of papers along with short descriptions and my personal contribution to each paper.

Title	Kvik: three-tier data exploration tools for flexible analysis of genomic data in epidemiological studies
Authors	**Bjørn Fjukstad**, Karina Standahl Olsen, Mie Jareid, Eiliv Lund, and Lars Ailo Bongo
Description	The initial description of Kvik, and how we used it to implement Kvik Pathways, a web application for browsing biologicap pathway maps integrated with gene expression data from the Norwegian Women and Cancer (NOWAC) cohort.
Contribution	Designed, implemented, and deployed Kvik and Kvik Pathways. Evaluated the system and wrote the manuscript.
Publication date	15 March 2015
Publication venue	F1000
Citation	[8] B. Fjukstad, K. S. Olsen, M. Jareid, E. Lund, and L. A. Bongo, "Kvik: three-tier data exploration tools for flexible analysis of genomic data in epidemiological studies," *F1000Research*, vol. 4, 2015

Title	Building Applications For Interactive Data Exploration In Systems Biology.
Authors	**Bjørn Fjukstad**, Vanessa Dumeaux, Karina Standahl Olsen, Michael Hallett, Eiliv Lund, and Lars Ailo Bongo.
Description	Describes how we further developed the ideas from Paper 1 into an approach that we used to build the MIXT web application.
Contribution	Designed, implemented, and deployed Kvik and the MIXT web application. Evaluated the system and wrote the manuscript.
Publication date	20 August 2017.
Publication venue	The 8th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB) August 20–23, 2017.
Citation	[9] B. Fjukstad, V. Dumeaux, K. S. Olsen, E. Lund, M. Hallett, and L. A. Bongo, "Building applications for interactive data exploration in systems biology," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 556–561

Title	Interactions Between the Tumor and the Blood Systemic Response of Breast Cancer Patients
Authors	Vanessa Dumeaux, **Bjørn Fjukstad**, Hans E Fjosne, Jan-Ole Frantzen, Marit Muri Holmen, Enno Rodegerdts, Ellen Schlichting, Anne-Lise Børresen-Dale, Lars Ailo Bongo, Eiliv Lund, Michael Hallett.
Description	Describes the MIXT system which enables identification of genes and pathways in the primary tumor that are tightly linked to genes and pathways in the patient Systemic Response (SR).
Contribution	Designed, implemented, and deployed the MIXT web application. Contributed to write the manuscript.
Publication date	28 September 2017.
Publication venue	PLoS Computational Biology
Citation	[10] V. Dumeaux, B. Fjukstad, H. E. Fjosne, J.-O. Frantzen, M. M. Holmen, E. Rodegerdts, E. Schlichting, A.-L. Børresen-Dale, L. A. Bongo, E. Lund *et al.*, "Interactions between the tumor and the blood systemic response of breast cancer patients," *PLoS Computational Biology*, vol. 13, no. 9, p. e1005680, 2017

Title	A Review of Scalable Bioinformatics Pipelines
Authors	**Bjørn Fjukstad**, Lars Ailo Bongo.
Description	This review survey several scalable bioinformatics pipelines and compare their design and their use of underlying frameworks and infrastructures.
Contribution	Wrote the manuscript.
Publication date	23 October 2017
Publication venue	Data Science and Engineering 2017.
Citation	[11] B. Fjukstad and L. A. Bongo, "A review of scalable bioinformatics pipelines," *Data Science and Engineering*, vol. 2, no. 3, pp. 245–251, 2017

Title	nsroot: Minimalist Process Isolation Tool Implemented With Linux Namespaces.
Authors	Inge Alexander Raknes, **Bjørn Fjukstad**, Lars Ailo Bongo.
Description	Describes a tool for process isolation built using Linux namespaces.
Contribution	Contributed to the manuscript, specifically to the literature review and related works.
Publication date	26 November 2017
Publication venue	Norsk Informatikkonferanse 2017.
Citation	[11] B. Fjukstad and L. A. Bongo, "A review of scalable bioinformatics pipelines," *Data Science and Engineering*, vol. 2, no. 3, pp. 245–251, 2017

Title	Transcription factor PAX6 as a novel prognostic factor and putative tumour suppressor in non-small cell lung cancer
Authors	Yury Kiselev, Sigve Andersen, Charles Johannessen, **Bjørn Fjukstad**, Karina Standahl Olsen, Helge Stenvold, Samer Al-Saad, Tom Dønnem, Elin Richardsen, Roy M Bremnes, and Lill-Tove Rasmussen Busund.
Description	This paper explores the possibility of using the PAX6 transcription factor as a prognostic marker in non-small cell lung cancer.
Contribution	Did the analyses to explore association between PAX6 gene expression and PAX6 target genes.
Publication date	22 March 2018
Publication venue	Scientific Reports 2018.
Citation	[12] Y. Kiselev, S. Andersen, C. Johannessen, B. Fjukstad, K. S. Olsen, H. Stenvold, S. Al-Saad, T. Donnem, E. Richardsen, R. M. Bremnes *et al.*, "Transcription factor pax6 as a novel prognostic factor and putative tumour suppressor in non-small cell lung cancer," *Scientific reports*, vol. 8, no. 1, p. 5059, 2018

Title	Reproducible Data Analysis Pipelines in Precision Medicine
Authors	**Bjørn Fjukstad**, Vanessa Dumeaux, Michael Hallett, Lars Ailo Bongo
Description	This paper outlines how we used the container centric development model to build walrus.
Contribution	Design, implementation and evaluation of walrus. Wrote the manuscript.
Publication date	TBA
Publication venue	TBA
Citation	[**?**]

1.6 Dissertation Plan

This thesis is organized as follows. Chapter 2 describes the characteristics of state-of-the-art biological datasets, the analysis required to extract knowledge from these, and the available tools and analysis frameworks. Chapter 3 describes in detail how we use a container centric development model to build a tool, walrus, to develop and execute deep analysis pipelines. In Chapter 4 we describe how we used the same model to develop applications to interactively explore results from statistical analyses. Finally, Chapter 5 concludes the work and discusses future directions.

2

Modern Biological Data Analysis

In this chapter we give a background in the different aspects of analyzing and exploring biological datasets. We highlight the necessary processing steps from data generation and all the way to interpretation of results.

3

Deep Analysis Pipelines

4

Interactive Exploration

We have also used the microservice architecture in an application where users can upload and explore air pollution data from Northern Norway.[13] In the project, air:bit, students from upper secondary schools in Norway collect air quality data from sensor kits that they have built and programmed. The web application lets the students upload data from their kits, and provides a graphical interface for them to explore data from their own, and other participating schools. The system consists of a web server frontend that retrieves air pollution data from a backend storage system to build interactive visualizations. It also integrates the data with other sources such as the Norwegian Institute for Air Research and the The Norwegian Meteorological Institute.

5

Conclusion

5.1 Lessons Learned

5.2 Broader Impact

5.3 Future Work

A

Publications

Bibliography

[1] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome biology*, vol. 11, no. 8, p. R86, 2010.

[2] O. Spjuth, E. Bongcam-Rudloff, G. C. Hernández, L. Forer, M. Giovacchini, R. V. Guimera, A. Kallio, E. Korpelainen, M. M. Kańduła, M. Krachunov *et al.*, "Experiences with workflows for automating data-intensive bioinformatics," *Biology direct*, vol. 10, no. 1, p. 43, 2015.

[3] Arvados, "Arvados | open source big data processing and bioinformatics," https://arvados.org, 2017, [Online; Accesssed: 16.08.2017].

[4] G. Kaushik, S. Ivkovic, J. Simonovic, N. Tijanic, B. Davis-Dusenbery, and D. Kural, "Rabix: an open-source workflow executor supporting recomputability and interoperability of workflow descriptions," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 22. NIH Public Access, 2016, p. 154.

[5] J. Vivian, A. A. Rao, F. A. Nothaft, C. Ketchum, J. Armstrong, A. Novak, J. Pfeil, J. Narkizian, A. D. Deran, A. Musselman-Brown *et al.*, "Toil enables reproducible, open source, big biomedical data analyses," *Nature Biotechnology*, vol. 35, no. 4, pp. 314–316, 2017.

[6] W. Tang, J. Wilkening, N. Desai, W. Gerlach, A. Wilke, and F. Meyer, "A scalable data analysis platform for metagenomics," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 21–26.

[7] Y. Diao, A. Roy, and T. Bloom, "Building highly-optimized, low-latency pipelines for genomic data analysis." in *CIDR*, 2015.

[8] B. Fjukstad, K. S. Olsen, M. Jareid, E. Lund, and L. A. Bongo, "Kvik: three-tier data exploration tools for flexible analysis of genomic data in epidemiological studies," *F1000Research*, vol. 4, 2015.

[9] B. Fjukstad, V. Dumeaux, K. S. Olsen, E. Lund, M. Hallett, and L. A. Bongo, "Building applications for interactive data exploration in systems biology," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 556–561.

[10] V. Dumeaux, B. Fjukstad, H. E. Fjosne, J.-O. Frantzen, M. M. Holmen, E. Rodegerdts, E. Schlichting, A.-L. Børresen-Dale, L. A. Bongo, E. Lund *et al.*, "Interactions between the tumor and the blood systemic response of breast cancer patients," *PLoS Computational Biology*, vol. 13, no. 9, p. e1005680, 2017.

[11] B. Fjukstad and L. A. Bongo, "A review of scalable bioinformatics pipelines," *Data Science and Engineering*, vol. 2, no. 3, pp. 245–251, 2017.

[12] Y. Kiselev, S. Andersen, C. Johannessen, B. Fjukstad, K. S. Olsen, H. Stenvold, S. Al-Saad, T. Donnem, E. Richardsen, R. M. Bremnes *et al.*, "Transcription factor pax6 as a novel prognostic factor and putative tumour suppressor in non-small cell lung cancer," *Scientific reports*, vol. 8, no. 1, p. 5059, 2018.

[13] B. Fjukstad, N. Angelvik, M. W. Hauglann, J. S. Knutsen, M. Grønnesby, H. Gunhildrud, and L. A. Bongo, "Low-cost programmable air quality sensor kits in science education," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. ACM, 2018, pp. 227–232.