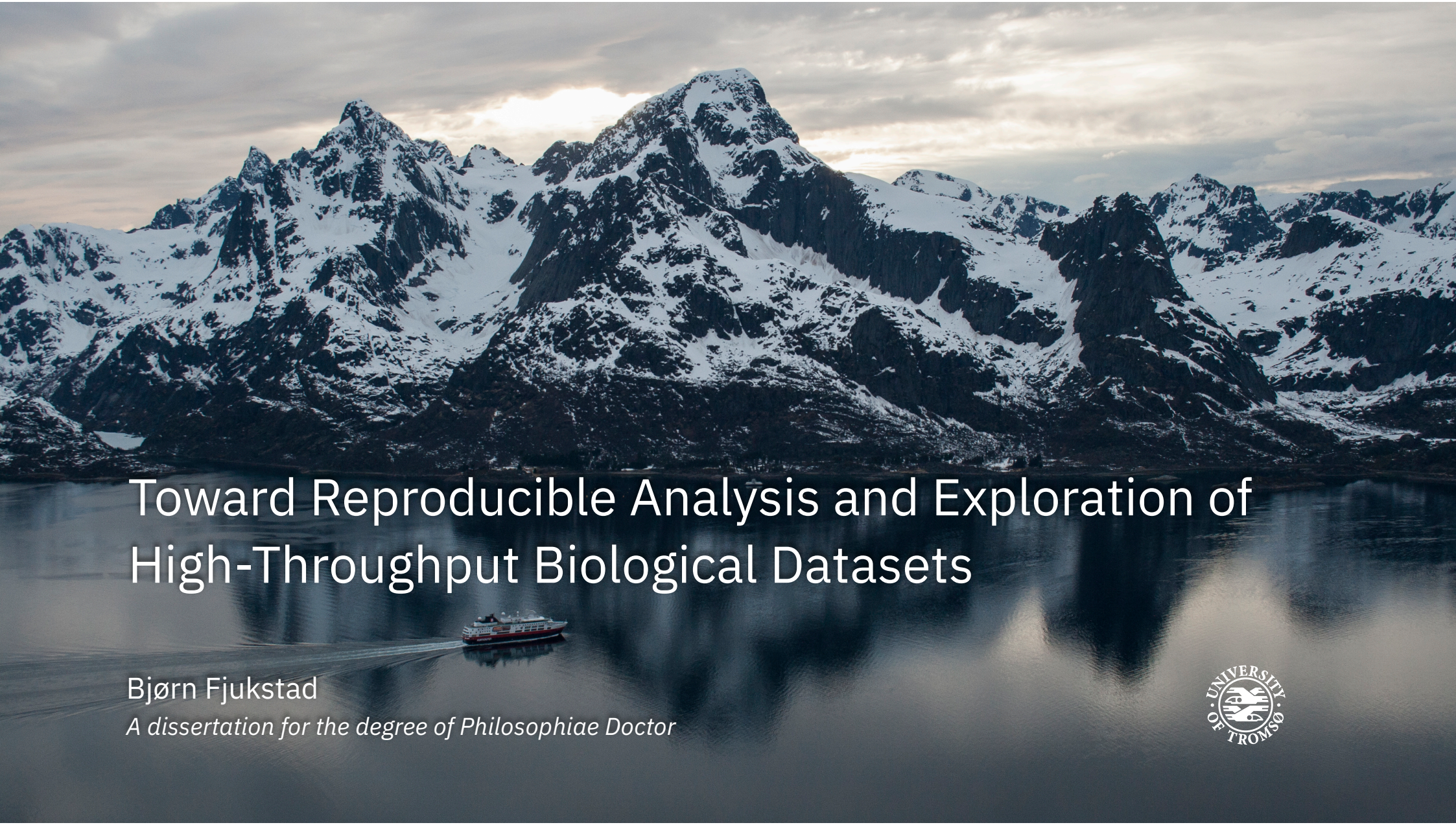


Toward Reproducible Analysis and Exploration of High-Throughput Biological Datasets

Bjørn Fjukstad

A dissertation for the degree of Philosophiae Doctor

UNIVERSITY OF TROMSØ

Great Potential

- High-throughput technologies and simpler access to datasets have revolutionized biology
- Datasets can reveal genetic basis of disease in patients, but require a collaborative effort
- Constantly ensure the quality of the data and analyses behind any interpretation
- Inaccurate results from improperly developed analyses can lead to negative consequences for patient care

Three different areas

- **Data management and analysis** for biological research datasets
- **Interactive data exploration applications** for use in epidemiological studies
- **Analysis pipelines** for complex genomic analysis in the clinic

	Data management and analysis	Interactive data exploration applications		Analysis pipelines
Application	**Pippeline**	**Kvik Pathways**	**MIxT**	**Clinical Sequencing Analysis**
Underlying System	**NOWAC R Package**	**Kvik**		**walrus**

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.

Challenges

- Store and track the necessary information to a data analyst when he or she is interpreting data, including code and data
- Integrate interactive applications with statistical analysis
- Integrate the different systems and tools needed by analysts
- Share and reuse applications and systems across research institutions

Contributions

- Simplified reproducing and reusing data and statistical analyses in the NOWAC study
- Developed interactive applications to enable novel insights in complex biological datasets
- Developed systems for reproducible analysis of genomic data in a clinical setting

Overview

- *Introduction*
- **Three focus areas**
  - **Data management and analysis**
  - **Interactive data exploration applications**
  - **Analysis pipelines**
- **Conclusion**

Data management and analysis

Different Datasets

- **High-Throughput Sequencing (HTS)** enable massively-parallel sequencing of DNA. These sequence millions of short base pairs, which are assembled together in the data analysis process.
- **Microarray technologies** enable us to study of actively expressed genes, or the transcriptome.
- **RNA-seq** technology based on HTS read millions of short base pairs in parallel, and can be used in gene expression analysis.
- **Surveys** are the traditional data collection method in epidemiology. Popular to integrate questionnaires with molecular data

Norwegian Women and Cancer

- Prospective population-based cohort that tracks 34% (170 000) of all Norwegian women born between 1943-57.
- The data collection started in NOWAC in 1991. Includes blood samples from 50.000 women, as well as more than 300 biopsies.
- First datasets generated in 2009. Now with additional miRNA, methylation, metabolomics, and RNA-seq datasets.

Previous practice

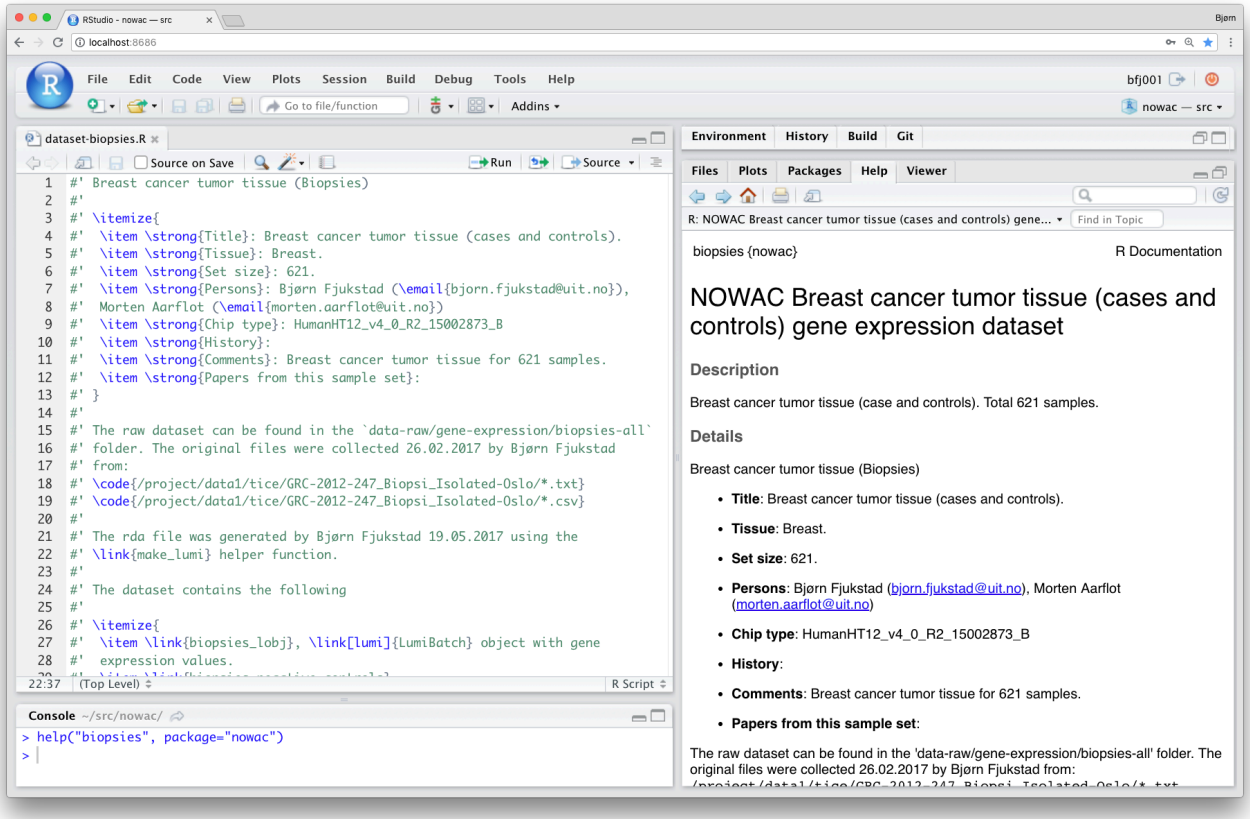
- Custom in-house storage for surveys and molecular datasets
- Preprocessing and data extraction through custom scripts by engineers, before statistical analyses by researchers.
- Cumbersome communication and data sharing through e-mail
- Statistical analysis using proprietary software

Challenges for Data Management and Analysis

- Keep track of available datasets, and their provenance.
- No standard storage platform or structure, and limited reports of the exported datasets in use in different research projects
- No standard approach to preprocess and initiate data analysis. Little practice of organized sharing of analysis code
- Difficult to reproduce due to lack of standardized preprocessing, sharing of analysis tools, and full documentation of the analyses

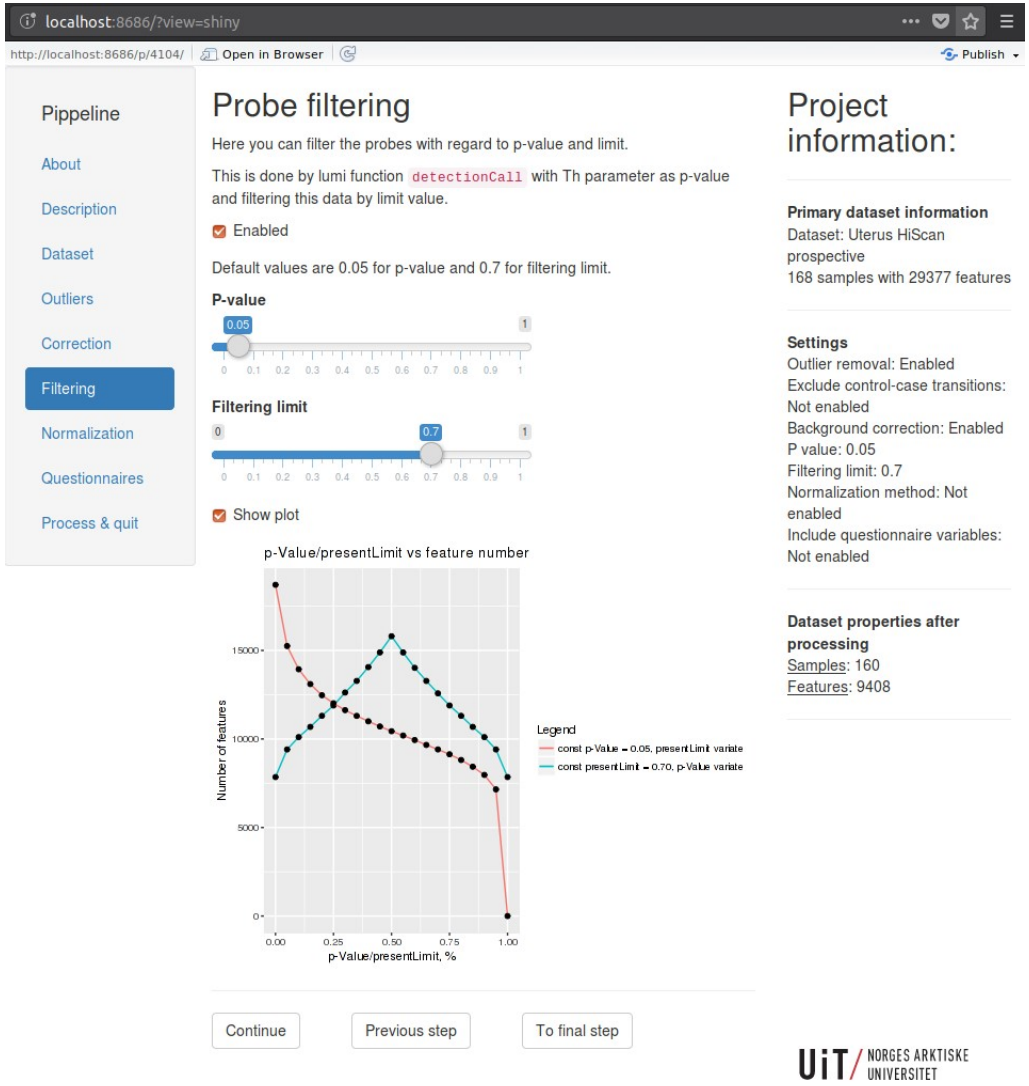
Solution

- The nowac R package: A software package in the R programming language that provides access to all data, documentation, and different utility functions for analysis.
- Pippeline: A data preparation system that uses the nowac R package to generate analysis-ready datasets for researchers
- Best practices for data analysis.



nowac R package

R package with data, documentation, and utility functions to analyze data from the NOWAC study



Pippeline

Standardized preprocessing of NOWAC datasets.

Best Practices

- Document every step in the analysis
- Generate reports and papers using code
- Version control everything
- Collaborate and share code through source code management systems

Conclusion

- A growing concern for the current level of reproducibility in science. With our approach we hope to facilitate for reproducible analyses
- Using R along with `git` are a potential drawbacks because of the users' required technical skill
- Like all software, it requires continuous maintenance
- The Pippeline now supports RNA-seq, Methylation, and microRNA datasets

Contributions

- Systematically organized and documented NOWAC datasets and analysis code
- Simplified reproducing and reusing data and statistical analyses in the NOWAC study

Overview

- *Introduction*
- *Three focus areas*
  - *Data management and analysis*
  - **Interactive data exploration applications**
  - **Analysis pipelines**
- **Conclusion**

Interactive applications to explore heterogeneous biological datasets

Interactive Data Exploration

- Visualization is central in both the analysis and understanding of biological functions
- Need specialized software to analyze and generate understandable visual representations of the complex datasets.
- Tools often integrate with online databases to allow researchers to study the data in the context of previous knowledge.

Use Case

- **High and Low Plasma Ratios of Essential Fatty Acids:** Visually explore gene expression measurements from healthy women with high and low plasma ratios of essential fatty acids.

- **Tumor-Blood Interactions in Breast Cancer Patients**: Explore genes and pathways in the primary tumor that are tightly linked to genes and pathways in the patient blood cells

Related Work

- Languages such as Python or R both provide a wealth of statistical packages and frameworks to analyze biological datasets
- Frameworks such as Shiny and OpenCPU allow application developers to build systems to interactively explore results from statistical analyses in R
- Tools such as Cytoscape or Circos support importing already analyzed datasets for visualization and exploration

Central Components

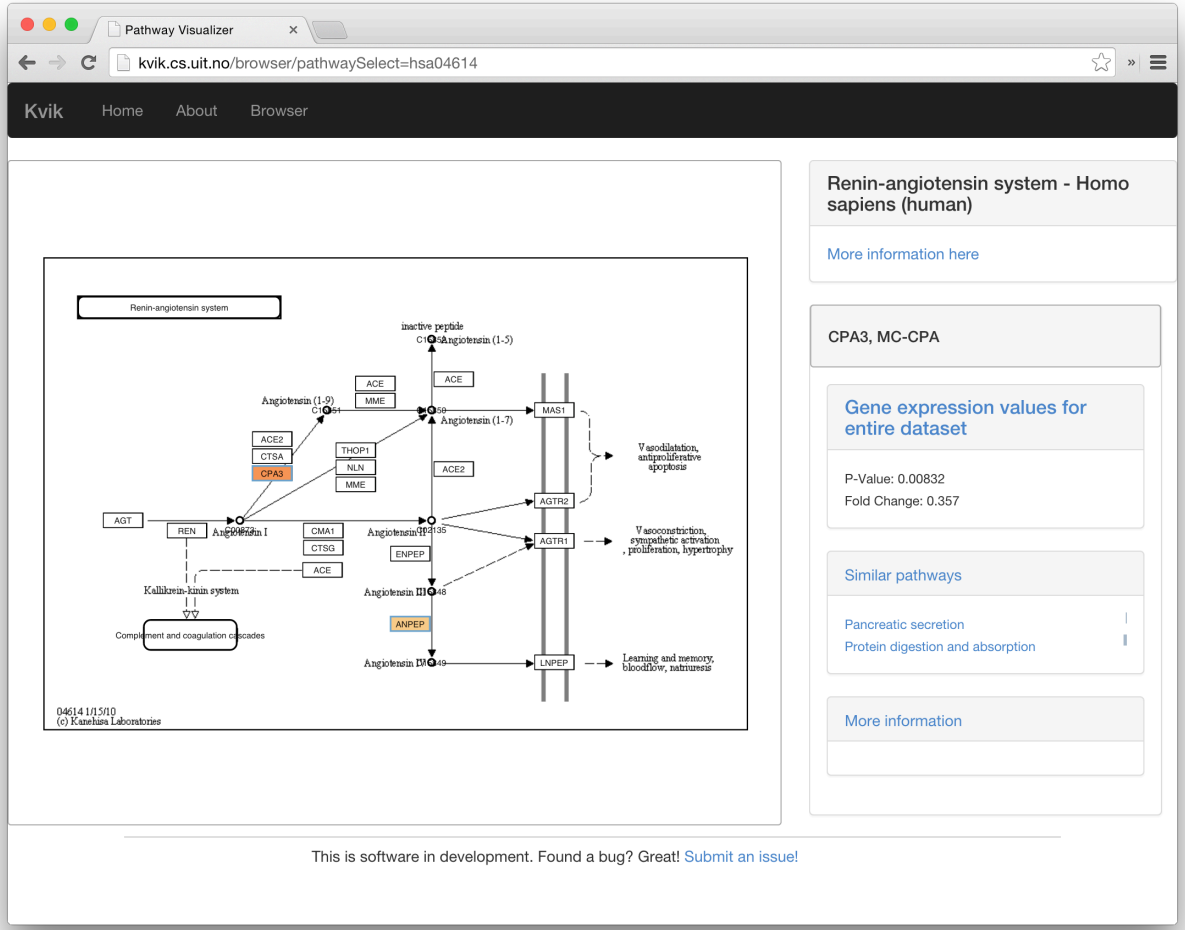
- A low-latency language-independent approach for integrating, or embedding, statistical software, such as R, directly in a data exploration application.

- A low-latency, language-independent interface to online reference databases in biology that users can query to explore results in context of results in context of known biology.

- A simple method for deploying and sharing the components of an application between projects.

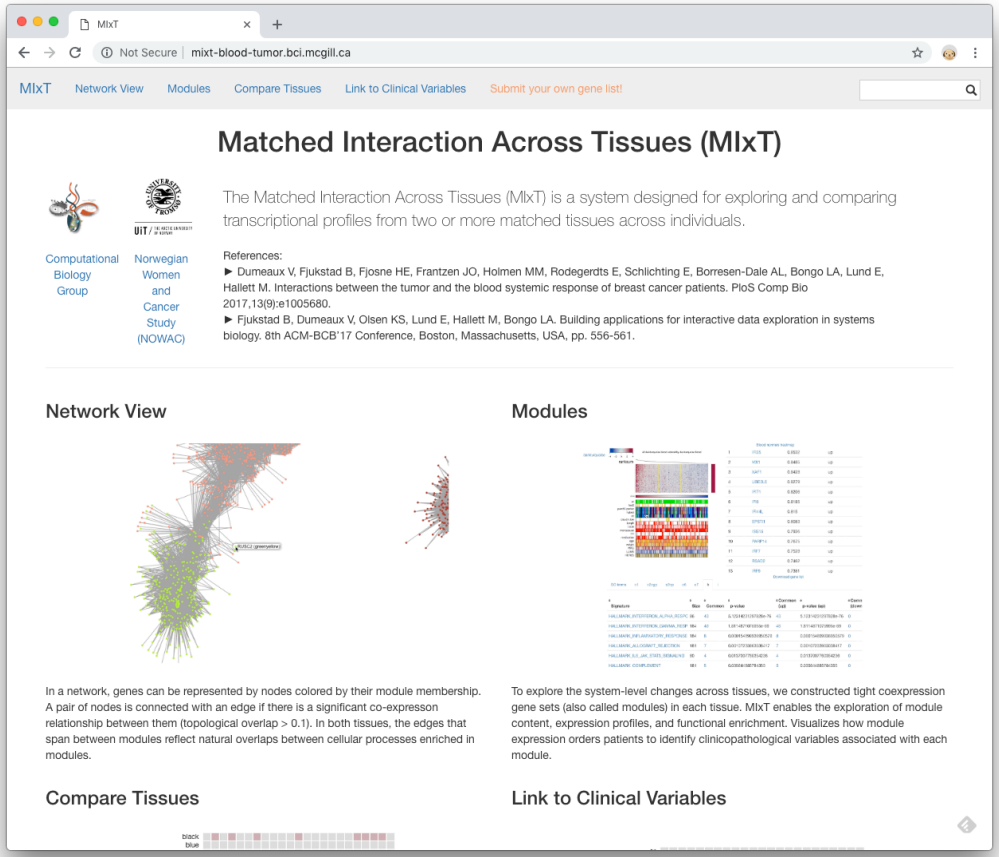
Kvik

- Collections of packages in the Go programming language that provides developers with the necessary tools to write interactive data exploration applications
- A compute service with an HTTP interface to the R programming language
- A database service that provides an interface to online resources such as E-Utilities, MSigDB, HGNC, and KEGG



Kvik Pathways

Visually explore gene expression measurements from healthy women with high and low plasma ratios of essential fatty acids.



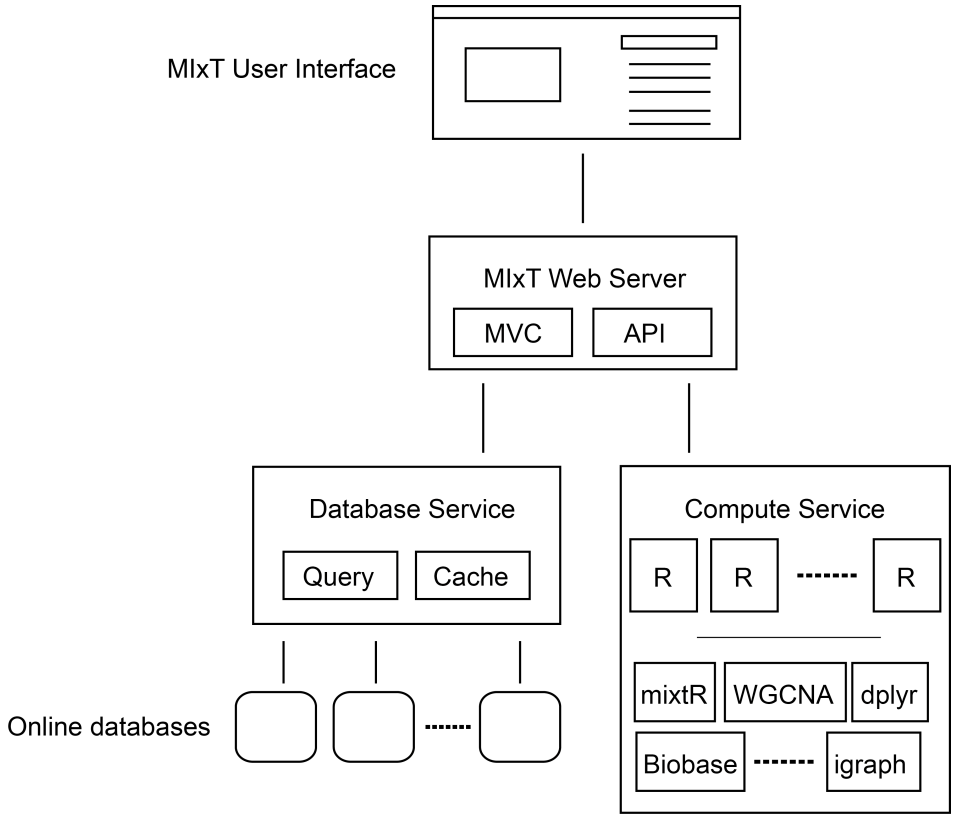
MIxT

Explore genes and pathways in the primary tumor that are tightly linked to genes and pathways in the patient blood cells

live demo

mixt-blood-tumor.bci.mcgill.ca

Design and Implementation



Evaluation

1. Investigate performance of R interface compared to similar service

2. Highlight the speedup of using a database service rather than the databases directly

Table 3.3: Time to complete the benchmark with different number of concurrent connections.

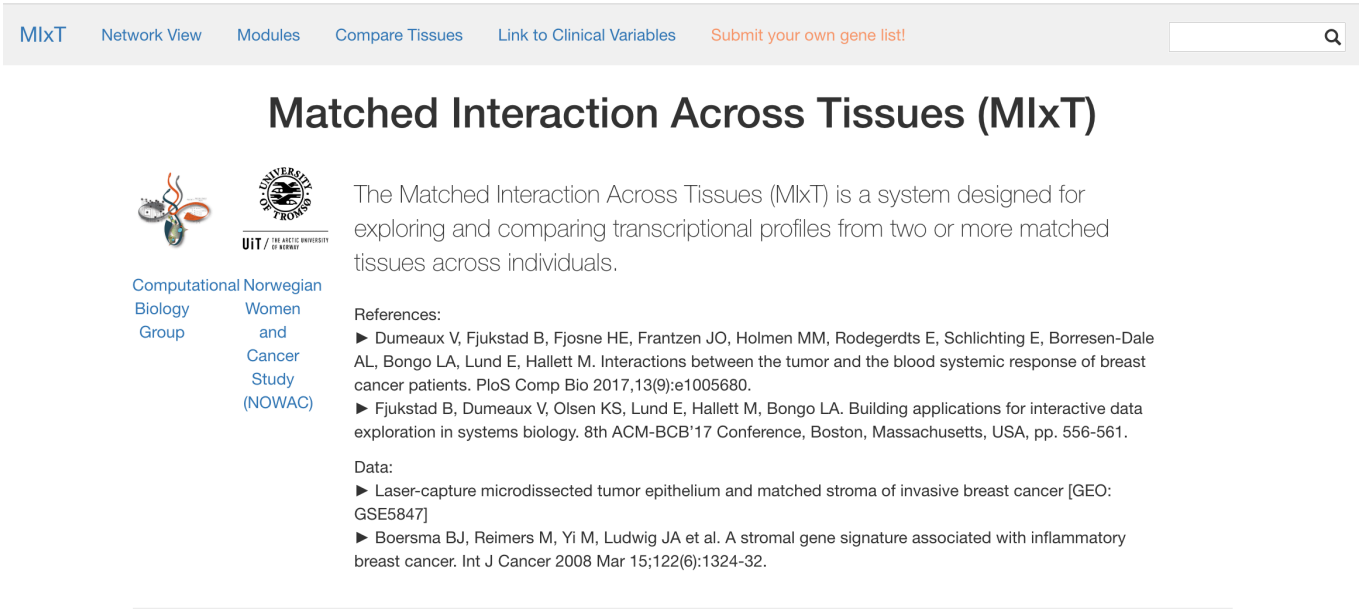
	1	2	5	10	15
Kvik	274ms	278ms	352ms	374ms	390ms
OpenCPU	500ms	635ms	984ms	1876ms	2700ms

Table 3.2: Time to retrieve a gene summary for a single gene, comparing different number of concurrent requests.

	1	2	5	10	15
No cache	956ms	1123ms	1499ms	2147ms	2958ms
Cache	64ms	64ms	130ms	137ms	154ms

Tumor Epithelium-Stroma Interactions in Breast Cancer

- Reuse of the MIxT web application with another dataset



Related Work

- A number of resources for exploring biological pathway maps: KEGG, KEGGViewer, Reactome, enRoute, and Pathview
- Few related systems to explore correlation networks similar to MIxT
- Common in the related applications is the decoupling of analysis and visualization
- Shiny and OpenCPU for developing applications that interface with R

Contributions

- Several interactive applications to enable novel insights in complex biological datasets
- Developed a system for integrating statistical analyses and biological databases with these applications
- Designed the applications to enable reuse and sharing of each of its components

Overview

- *Introduction*
- *Three focus areas*
  - *Data management and analysis*
  - *Interactive data exploration applications*
  - **Analysis pipelines**
- **Conclusion**

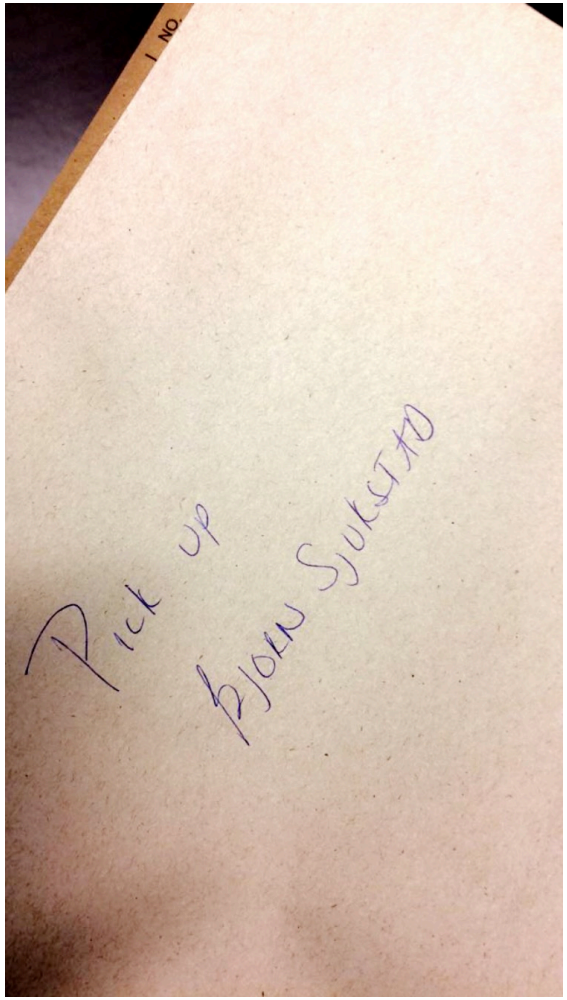
Long-running analysis pipelines for complex genomic analyses in the clinic

Precision Medicine

- Precision medicine uses patient-specific molecular information to diagnose and categorize disease to tailor treatment to improve health outcome.
- High-throughput sequencing is currently the main technology to facilitate personalized diagnosis and treatment
- Downstream computational analysis and interpretation is the major cost. $1000 genome, $1,000,000 analysis

Genomic Analysis

- Analysis pipelines consist of many steps that transform raw data into interpretable results
- It is necessary to carefully explore different tools and parameters to answer a dedicated question
- Improperly developed analysis pipelines may generate inaccurate results, which may have negative consequences for patient care



Use Case

- **Provide an extensive comparison of the metastasis against the primary, and to identify the molecular drivers of the tumor**
- Previous: Identify the molecular signature of the patient's tumor and germline with DNA sequence data from the patient's primary tumor and adjacent normal cells

14 October 2016, pick up of biological sample at McGill University Health Centre

Lessons from the initial analysis pipeline

- The pipeline tools should be kept for later use, either as binaries or code
- Datasets and databases should be stored along with the pipeline description
- It should be easy to install and include new tools to a new or existing pipeline

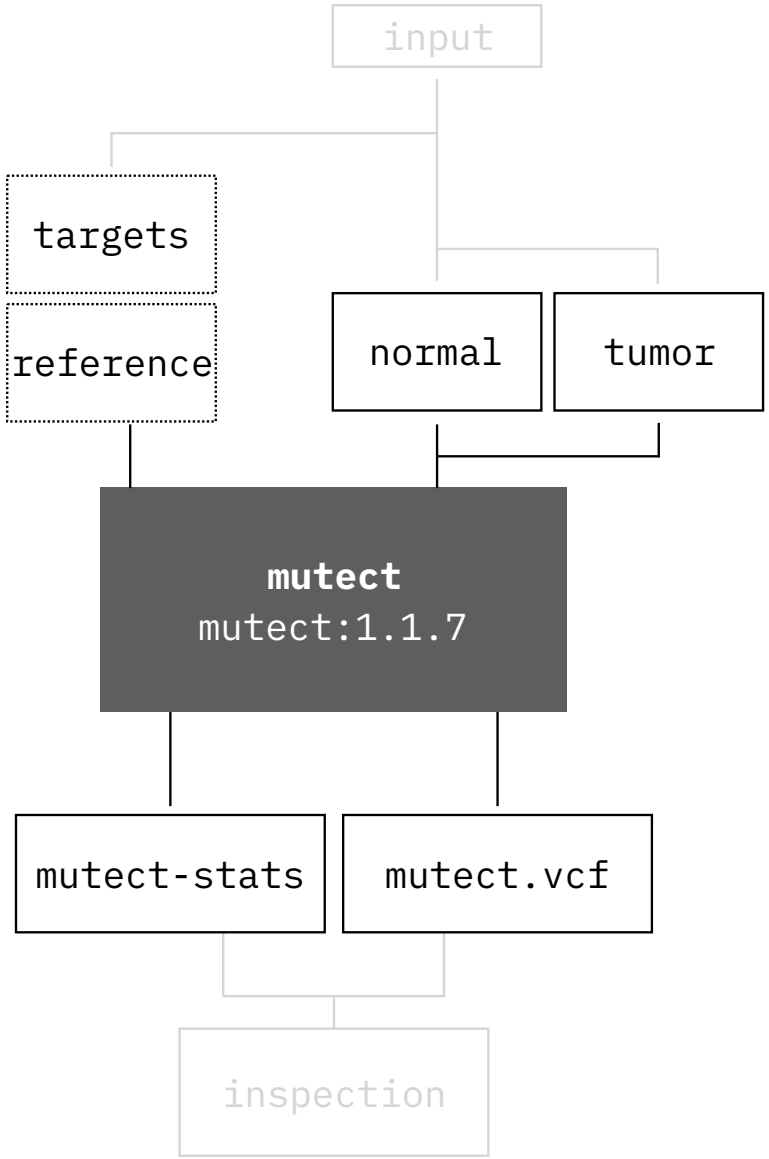
Related Work

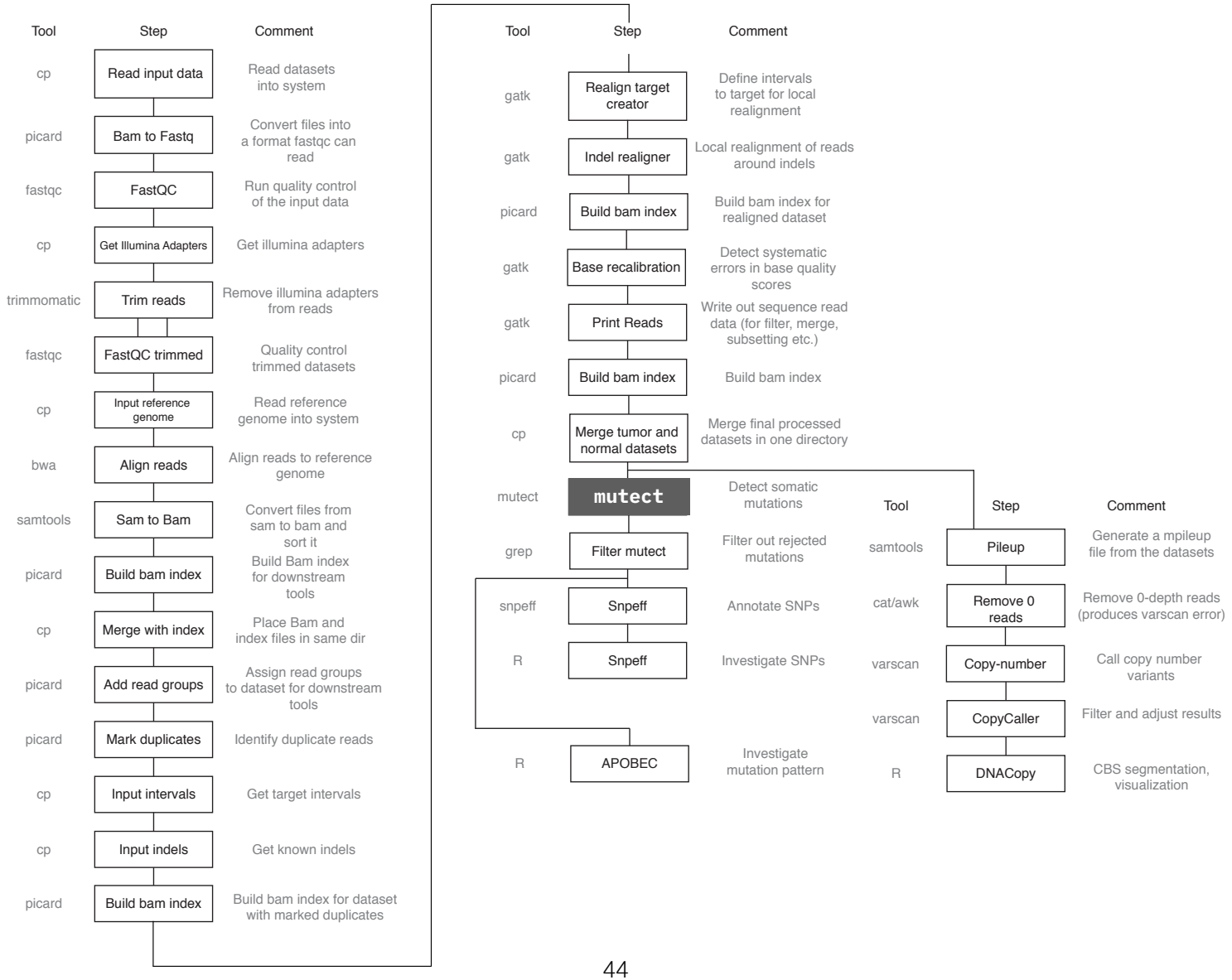
- A wealth of pipeline description formats and workflow managers available
- Common Workflow Language (CWL) is a popular specification formation that is supported by many workflow managers, e.g. Galaxy, Rabix, Toil, and AWE. Implementation specific data management.
- Pachyderm provides version control for data, and reproducible data analysis pipelines. Runs on top of Kubernetes.
- Snakemake is a popular tool that has lately got support software containers

walrus

- Tool for developing and executing data analysis pipelines
- Stores: tool versions, tool parameters, input data, intermediate data, output data, along with the execution environment
- Users write pipeline descriptions in JSON/YAML files
- Users use the commandline interface which executes the pipeline

```
{
    "Name": "mutect",
    "Image": "fjukstad/mutect:1.1.7",
    "Cmd": [
        "--analysis_type","MuTect",
        "--reference_sequence","/walrus/input/reference.fasta",
        "--input_file:normal","/walrus/input/normal.bam",
        "--input_file:tumor","/walrus/input/tumor.bam",
        "-L", "/walrus/input/targets.bed",
        "--out", "/walrus/mutect/mutect-stats-txt",
        "--vcf","/walrus/mutect/mutect.vcf"
    ],
    "Inputs":[
      "input"
    ]
}
```




Tool
Step
Comment
cp
Read input data
Read datasets into system
picard
Bam to Fastq
Convert files into a format fastqc can read
fastqc
FastQC
Run quality control of the input data
cp
Get Illumina Adapters
Get illumina adapters
trimmomatic
Trim reads
Remove illumina adapters from reads
fastqc
FastQC trimmed
Quality control trimmed datasets
cp
Input reference genome
Read reference genome into system
bwa
Align reads
Align reads to reference genome
samtools
Sam to Bam
Convert files from sam to bam and sort it
picard
Build bam index
Build Bam index for downstream tools
cp
Merge with index
Place Bam and index files in same dir
picard
Add read groups
Assign read groups to dataset for downstream tools
picard
Mark duplicates
Identify duplicate reads
cp
Input intervals
Get target intervals
cp
Input indels
Get known indels
picard
Build bam index
Build bam index for dataset with marked duplicates
Tool
Step
Comment
gatk
Realign target creator
Define intervals to target for local realignment
gatk
Indel realigner
Local realignment of reads around indels
picard
Build bam index
Build bam index for realigned dataset
gatk
Base recalibration
Detect systematic errors in base quality scores
gatk
Print Reads
Write out sequence read data (for filter, merge, subsetting etc.)
picard
Build bam index
Build bam index
cp
Merge tumor and normal datasets
Merge final processed datasets in one directory
mutect
mutect
Detect somatic mutations
grep
Filter mutect
Filter out rejected mutations
snpeff
Snpeff
Annotate SNPs
R
Snpeff
Investigate SNPs
R
APOBEC
Investigate mutation pattern
Tool
Step
Comment
samtools
Pileup
Generate a mpileup file from the datasets
cat/awk
Remove 0 reads
Remove 0-depth reads (produces varscan error)
varscan
Copy-number
Call copy number variants
varscan
CopyCaller
Filter and adjust results
R
DNACopy
CBS segmentation, visualization

Pipeline Execution

- All tools are Docker images, and each pipeline step is run within a Docker container. Simplifies dependency management
- Users can create these images themselves, or use existing images from e.g. BioContainers
- A pipeline is run on a single server, but supports data and task parallelism

Data Management

- We version control all data: input, intermediate, and output
- We use `git` for the pipeline description, and `git-lfs` for the datasets
- The commandline interface takes care of tracking, and optionally restoring previous datasets

Results

- Has shown its usability in a clinical setting. We analyzed a patient’s metastasis to discover Single Nucleotide Polymorphisms(SNPs), genomic variants and somatic mutations

- Implemented a variant calling pipeline for a public dataset to evaluate performance and resource usage.

Discussion

- Still no standard for representing or sharing workflows. CWL is gaining popularity and more systems now support it
- In walrus we propose a solution where researchers can share pipeline descriptions *and* datasets
- Forcing users to package tools in Docker images can help ensure reproducibility

Contributions

- Developed systems for reproducible analysis of genomic data in a clinical setting
- Analyzed a patient’s metastasis to discover SNPs, genomic variants and somatic mutations

Overview

- *Introduction*
- *Three focus areas*
  - *Data management and analysis*
  - *Interactive data exploration applications*
  - *Analysis pipelines*
- **Conclusion**

Conclusion

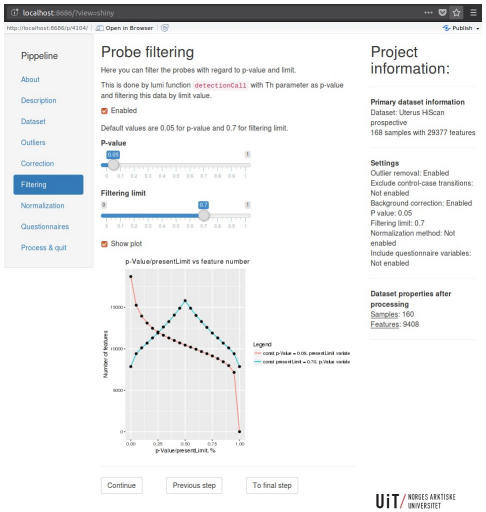
Lessons Learned and Future Research Challenges

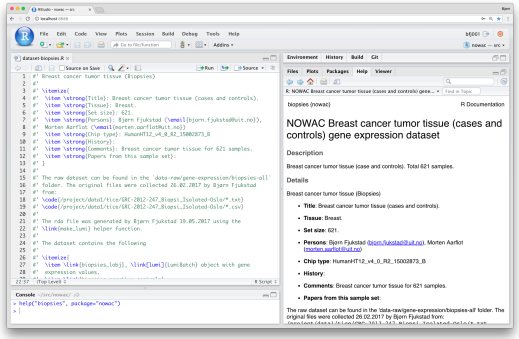
- Systems require the integration of different tools and programming languages
- Further investigate versioning and sharing of large biological datasets
- Distributed orchestration and execution of biological pipelines using tools such as Kubernetes
- Our applications and tools require continuous maintenance and care

Conclusion

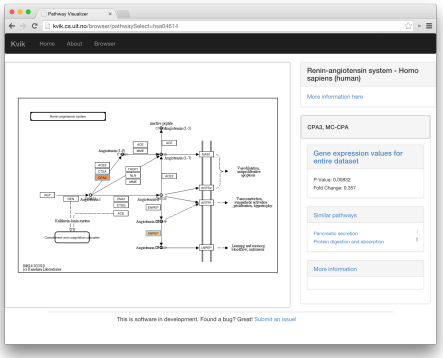
- We have proposed a solution for storing, managing, analyzing, and exploring biological datasets that facilitates sharing, reuse, and reproducibility.
- We have shown the viability of the approach through real-world applications in systems epidemiology and precision medicine
- Datasets and computing systems will evolve, and we believe our approach can offer a new perspective on developing applications for exploring and analyzing data.

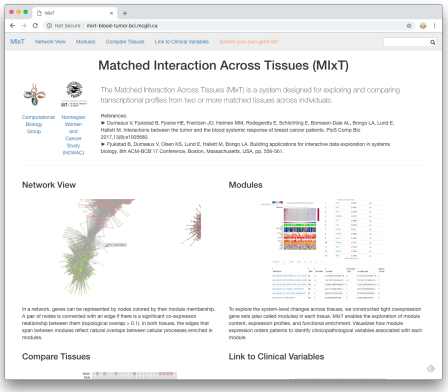
Data management and analysis



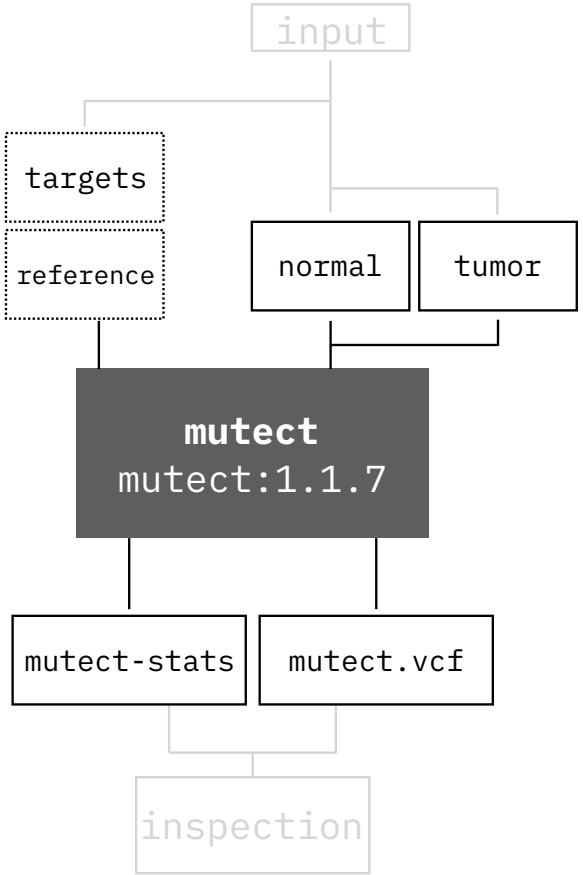


Interactive data exploration applications





Analysis pipelines



Contributions

- Simplified reproducing and reusing data and statistical analyses in the NOWAC study
- Developed interactive applications to enable novel insights in complex biological datasets
- Developed systems for reproducible analysis of genomic data in a clinical setting

Papers

1. Kvik: three-tier data exploration tools for flexible analysis of genomic data in epidemiological studies
   ***Bjørn Fjukstad**, Karina Standahl Olsen, Mie Jareid, Eiliv Lund, and Lars Ailo Bongo*

2. Building Applications For Interactive Data Exploration In Systems Biology
   ***Bjørn Fjukstad**, Vanessa Dumeaux, Karina Standahl Olsen, Michael Hallett, Eiliv Lund, and Lars Ailo Bongo*

3. Interactions Between the Tumor and the Blood Systemic Response of Breast Cancer Patients
   *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, and Michael Hallett*

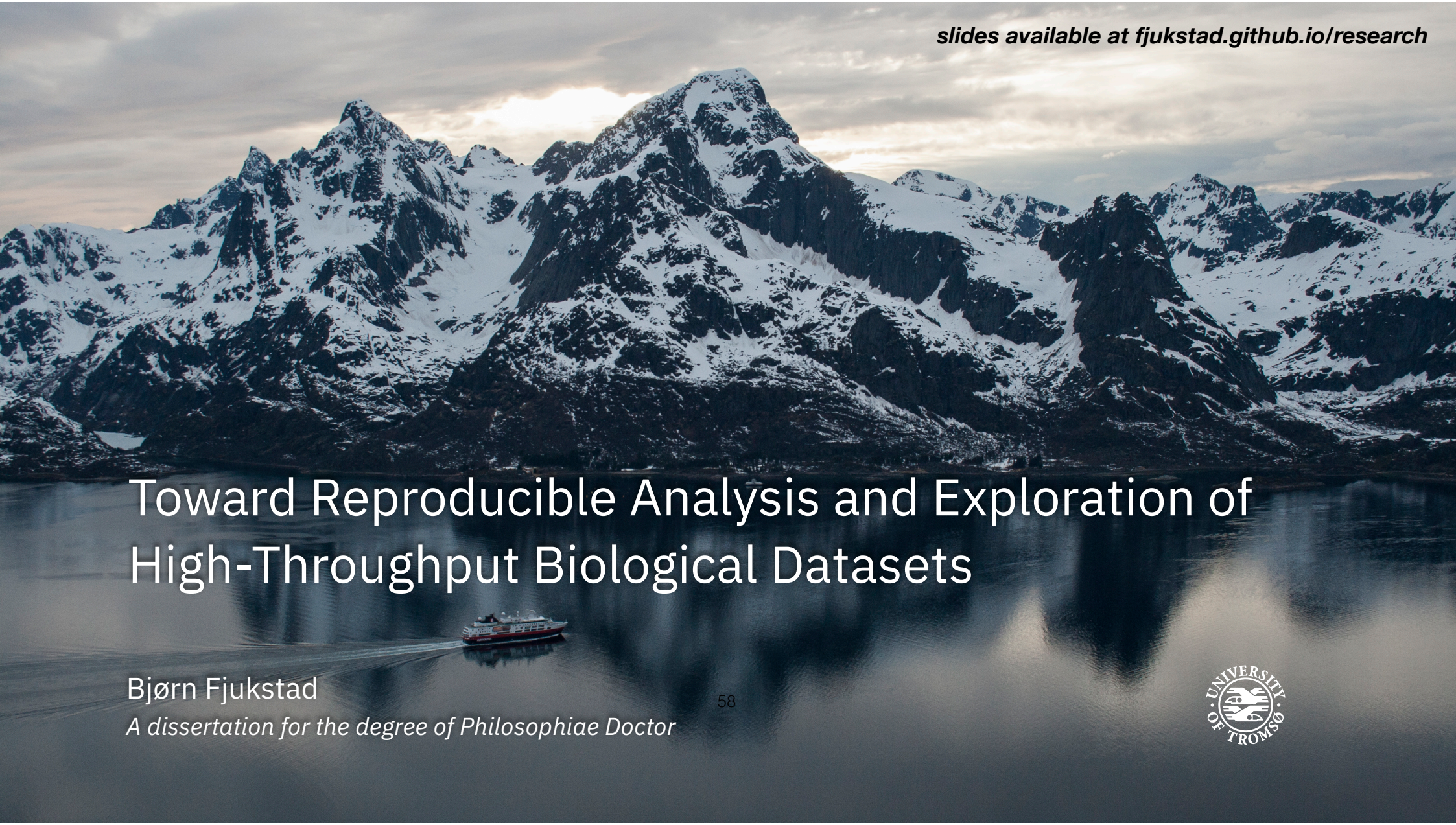
4. A Review of Scalable Bioinformatics Pipelines
   ***Bjørn Fjukstad** and Lars Ailo Bongo*

5. nsroot: Minimalist Process Isolation Tool Implemented With Linux Namespaces
   *Inge Alexander Raknes, **Bjørn Fjukstad**, and Lars Ailo Bongo.*

6. Reproducible Data Analysis Pipelines for Precision Medicine
   ***Bjørn Fjukstad**, Vanessa Dumeaux, Michael Hallett, and Lars Ailo Bongo*

Acknowledgement

- Advisors: **Lars Ailo Bongo**, **Karina Standahl Olsen**, and **Eiliv Lund**
- Biological Data Processing Systems Lab: **Einar Holsbø**, **Morten Grønnesby**, **Edvard Pedersen, Nikita Shvetsov, Jo Inge Arnes, Tengel Skar**
- Norwegian Women and Cancer (NOWAC)
- Concordia University: **Vanessa Dumeaux** and **Mike Hallett**
- Department of Computer Science at UiT



Toward Reproducible Analysis and Exploration of High-Throughput Biological Datasets

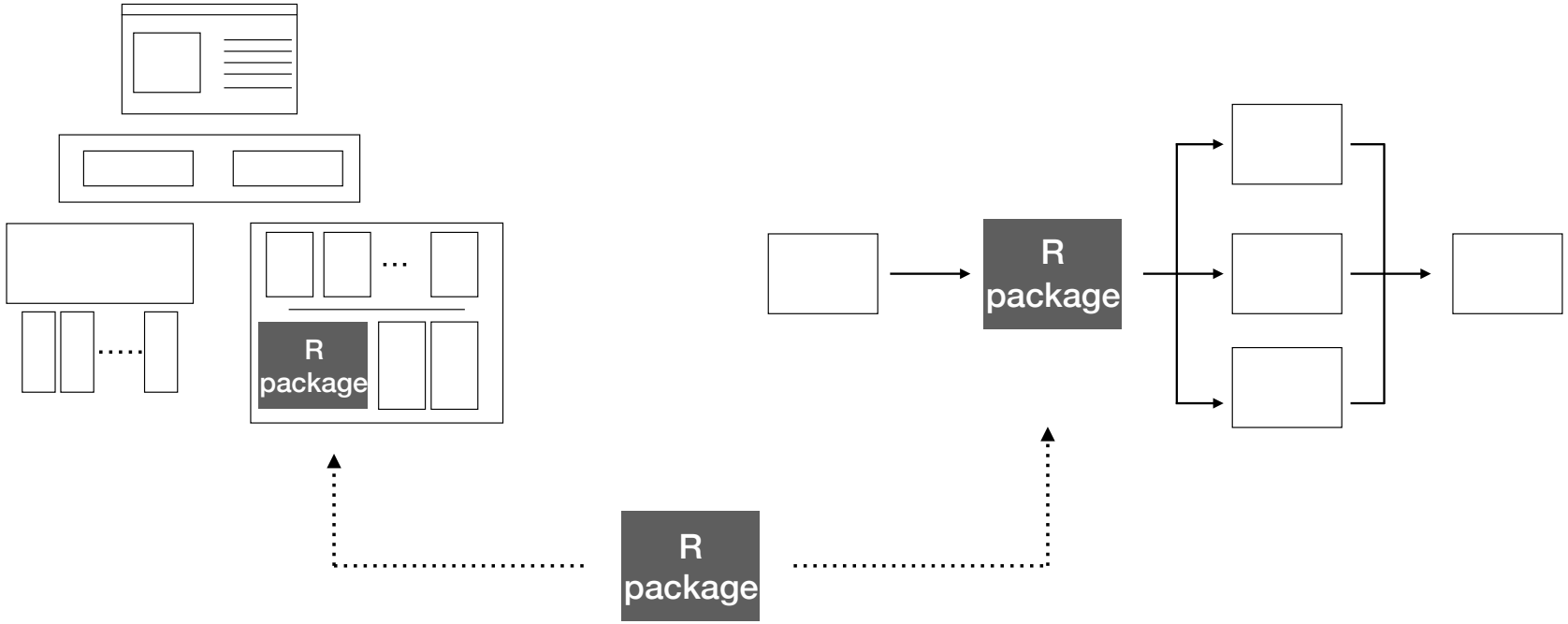
Bjørn Fjukstad

A dissertation for the degree of Philosophiae Doctor

UNIVERSITY OF TROMSØ

Data exploration application

Analysis pipeline



Data management and analysis

Table 4.1: Runtime and storage use of the example variant-calling pipeline developed with `walrus`.

Experiment	Task	Runtime	Storage Use
1	Run pipeline with default configuration	21 hours 50 minutes	235 GB
2	Run the default pipeline with version control of data	23 hours 9 minutes	470 GB
3	Re-run the pipeline with modified indel realignment parameter	13 hours	500 GB
4	Restoring pipeline back to the default configuration	$<$ 1 second	500GB