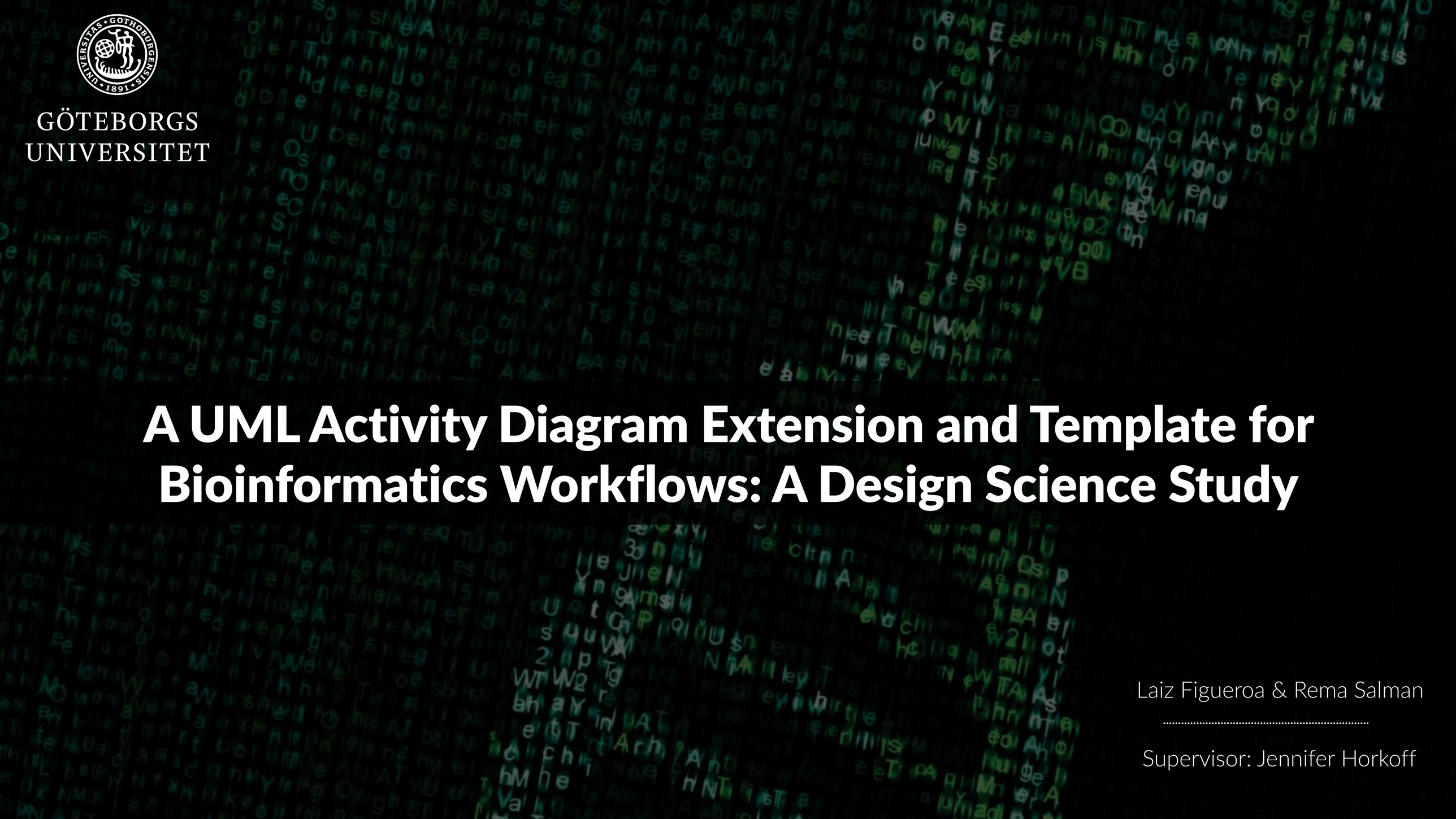


GÖTEBORGS UNIVERSITET

A UML Activity Diagram Extension and Template for Bioinformatics Workflows: A Design Science Study

Laiz Figueroa & Rema Salman

Supervisor: Jennifer Horkoff

GÖTEBORGS UNIVERSITET

Introduction

Bioinformatics

- Biology and computational methods together [1]
- Uses several tools to generate data
- Tools' connections are represented by workflows (pipelines)

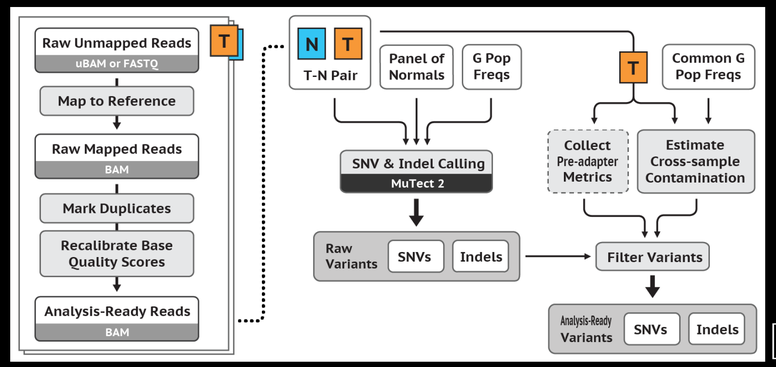
Workflow & Pipeline

- Sequence of tasks from initialisation to producing final results [2]
- Shepherding files through a series of transformations [3]

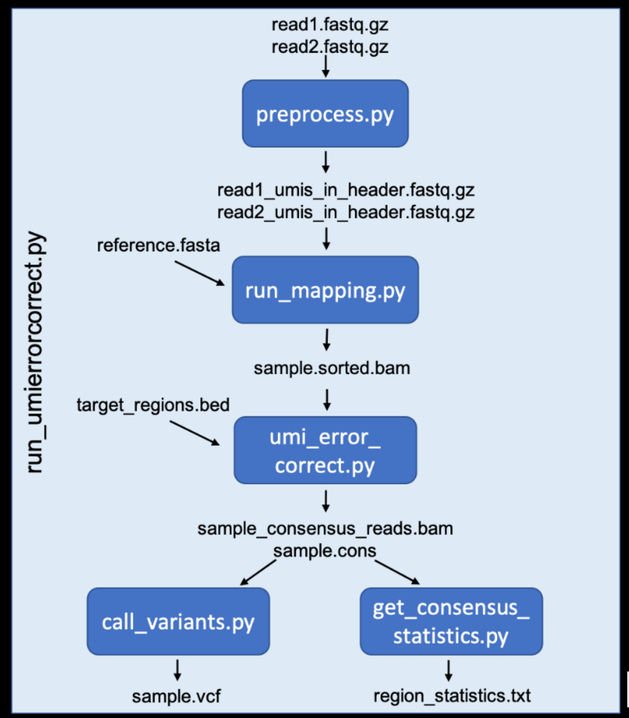
Usage

- These workflows need to be followed precisely to generate the correct data [4]

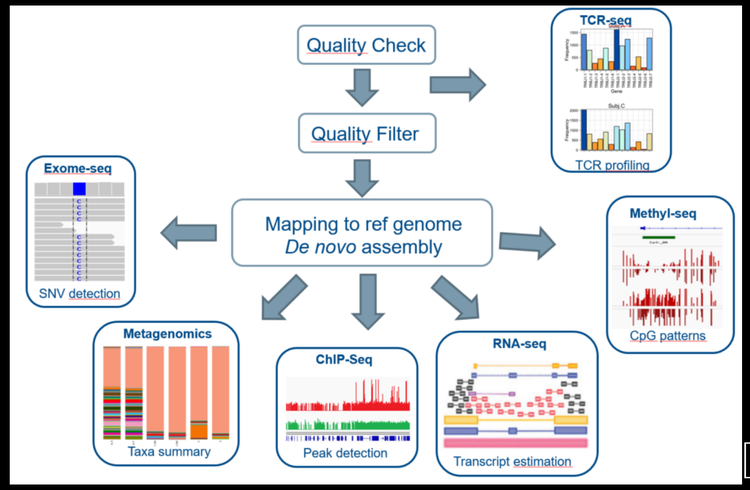
Problem



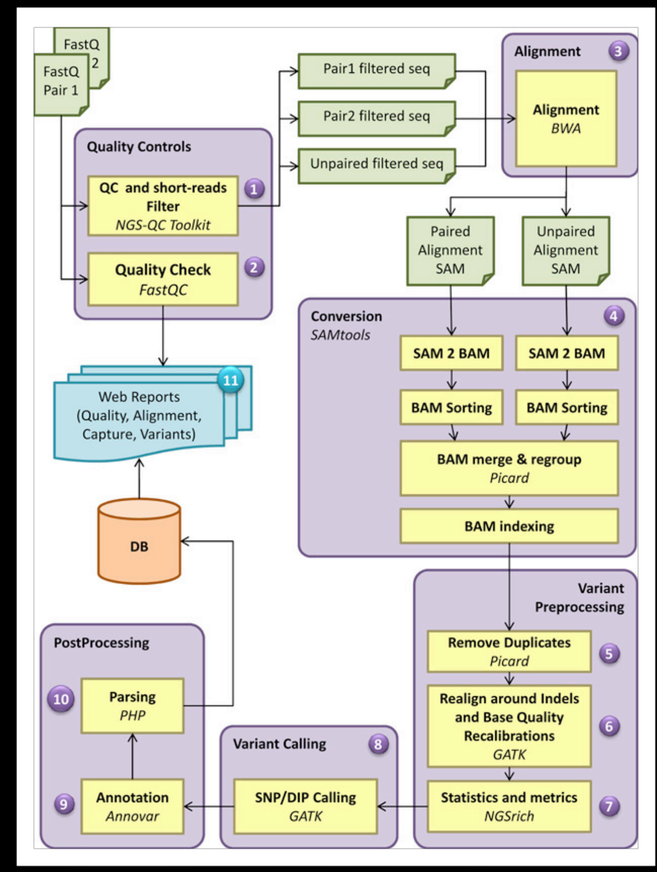
[9]



[11]



[11]



[10]

Quality assessment of the sequence reads was performed by generating QC statistics with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Read alignment to the reference human genome (hg19,UCSC assembly, February 2009) was done using BWA (1) with default parameters. [A summary of the sequencing data is shown in Table X.] After removal of PCR duplicates (Picard tools, http://picard.sourceforge.net) and file conversion (samtools (2)) quality score recalibration, indel realignment and variant calling were performed with the GATK package(3). Variants were annotated with Annovar (4) using a wide range of databases such as dbSNP build 135 (5), dbNSFP (6), KEGG (7), the Gene Ontology project (8), MITOMAP (9) and tracks from the UCSC. [11]

GÖTEBORGS UNIVERSITET

Background

Horkoff et al. [8]

- Used several modelling languages
  - UML activity diagram most suitable
- Identified concepts gaps
  - Motivations
  - Sources
  - Thresholds
  - Files
- Suggested further study to extend the language
- Proposed a draft for workflow elicitation

GÖTEBORGS UNIVERSITET

Research Question

How can we extend the UML activity diagram and use a template for workflow documentation to understand and improve bioinformatics workflows?

GÖTEBORGS UNIVERSITET

Research Purpose

Extend the UML AD meta-model, create its new concrete syntax, and generate a Workflow Documentation Specification Template (WDST)

- Increase efficiency to manage workflows
- Establish a shared understanding and consistency between the activities
- Create a sharable documentation set
- Provide a way to train new bioinformaticians
- Identify problems in workflows

Facilities & Sample

The head of Bioinformatics Core Facility

Purposive sampling technique

CRITERIA

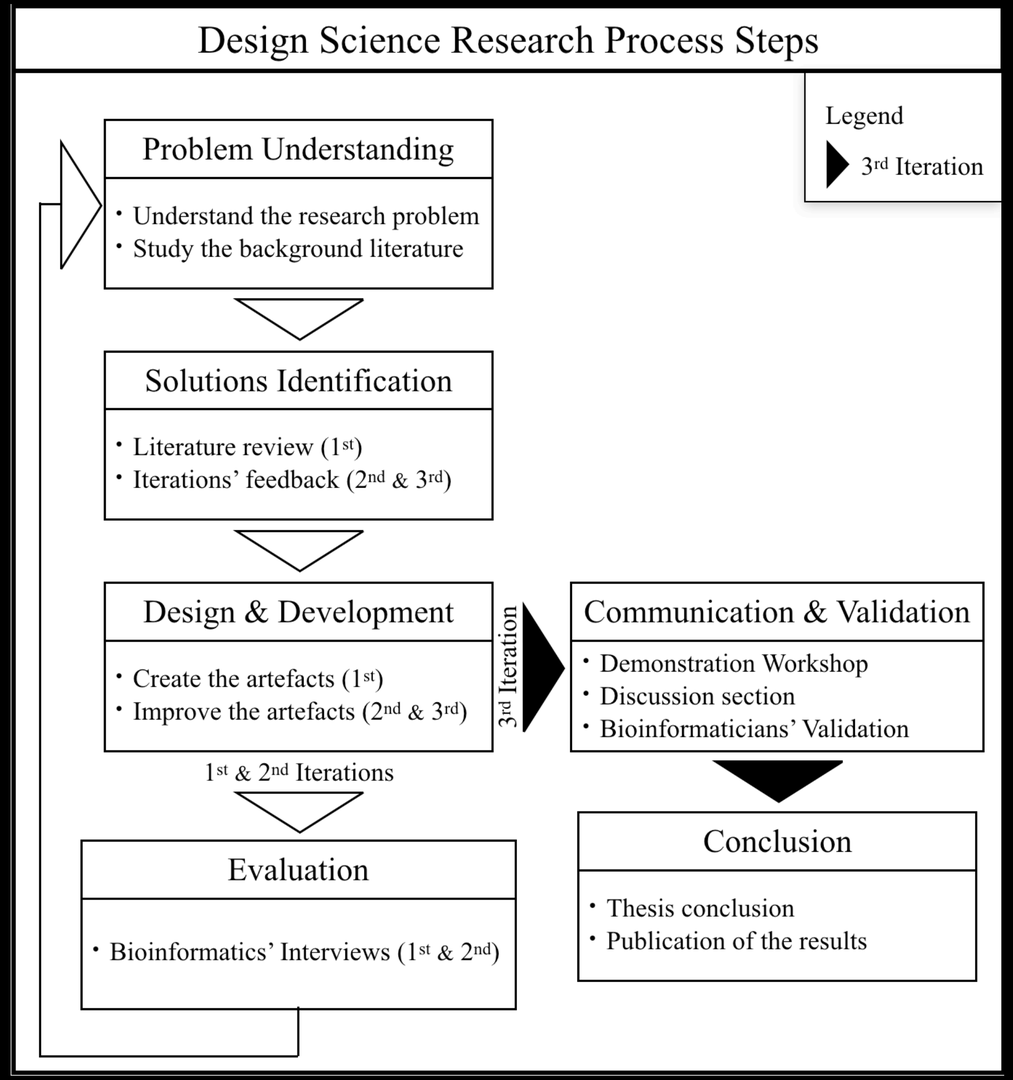
Bioinformaticians with workflows' knowledge

6

- Bioinformatics Core Facility
- Genomic Medicine Sweden
- Translational Genomics Platform

GÖTEBORGS UNIVERSITET

Methodology



1st

Recorded semi-structured interview
5 bioinformaticians
Transcript using *Temi*
Thematic analysis

2nd

Recorded semi-structured interview intercalated with artefacts' test
5 bioinformaticians - 1 new
Think aloud protocol - log
Transcript using *Temi*
Thematic analysis

3rd

Recorded workshop discussion
6 bioinformaticians - 1 new
Validation questions using *Mentimeter*
Transcript using *Temi*
Thematic analysis
Suggest further studies

UML Activity Diagram Extension Meta-model

RQ 1.1

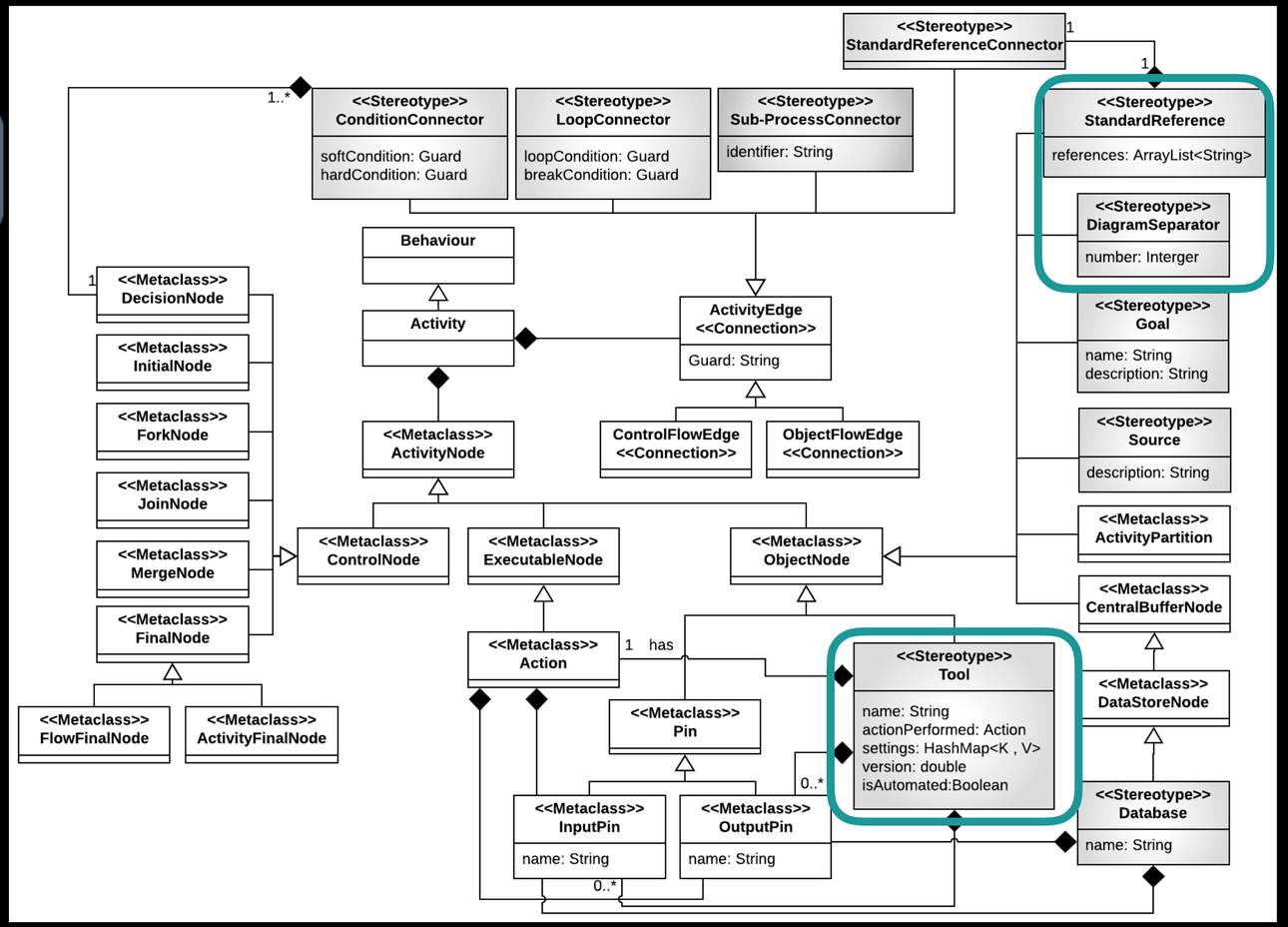
What are the defining and unique characteristics of bioinformatics workflows compared to standard workflows?

9 highly used characteristics

3 considered unique

6 bridge between standard workflow and UML AD

Added **data flow** behaviour to AD

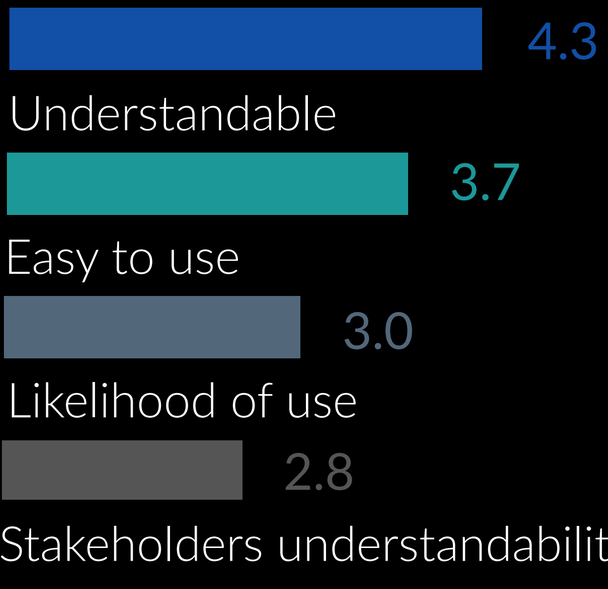


GÖTEBORGS UNIVERSITET

Concrete Syntax

RQ 1.2

How should workflows, including the concepts discovered in RQ1.1 be visualised to be understandable by the bioinformaticians?

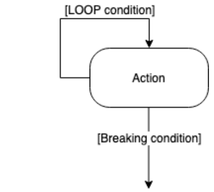
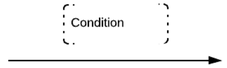
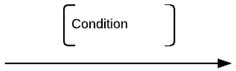
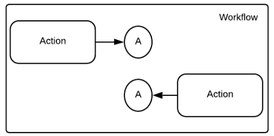
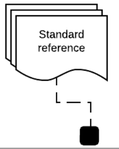
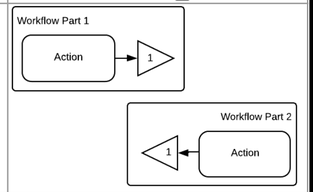


Use the

concrete syntax

with

labels

Name	Base Class	Description	Notation
Loop	ActivityEdge	An iterative set of activities and actions represents until reaching the defined condition.	
SoftCondition	ActivityEdge	Represent an outcome of a test based on a condition with a limited soft-threshold value. The condition is predefined guards on the outgoing edges.	
HardCondition	ActivityEdge	Represent an outcome of a test based on a condition with a limited hard-threshold value. The condition is predefined guards on the outgoing edges.	
Sub-processConnector	ActivityEdge	Used to connect the sub-processes parts within the same diagram.	
StandardReferenceConnector	Activity Edge	A connector used between the dark input and the multiple documents notations to represent the standard reference.	
StandardReference	ObjectNode	Data that is used to make comparison. This data is normally standards followed. For example, human genome.	
DiagramSeparator	ObjectNode	A labeled triangle that represents the connection point with an other part of the diagram from other page.	
Source	ObjectNode	A link, document title, person's name which are the source or responsible for a specific set of actions.	
Tool	ObjectNode	A tool or software used to perform an activity with a description of the activity. That is automated operated.	
	ObjectNode	A tool or software used to perform an activity with a description of the activity. That is manually operated.	
Database	DataStoreNode	A structured set of data that is accessible in various ways.	

GÖTEBORGS UNIVERSITET

WDST

RQ 1.3

How can we design a useful and understandable template to document the concepts from RQ1.1 from the bioinformaticians viewpoint?

2.0
Understandable

1.7
Easy to use

1.3
Likelihood of use

1
Stakeholders understandability

Unanimously
disliked
failed attempt

- Automatically generate documentation after the workflow is drawn
- Must contain the tools section
- The amount of text and technicality should be as low as possible

Workflow Description Specification

Workflow ID:		<<the workflow name or identifier>>
Date of creation:	<<date in which this document was created>>	Number of steps: <<amount of steps>>
Workflow version: <<version of this document>>	Modification date: <<date of modification>>	Workflow creator: <<name>>

Workflow

Workflow goal:	<<what do you want to achieve with this workflow?>>
Workflow source:	<< Is this workflow created locally? or it follows a reference - in that case, add link to the reference or name the person>>
Workflow responsible:	<<person who signs the final output or who uses this workflow>>

First Step (Start point)	Final Step (End point)
Step ID: <<The name or identifier of the start step>>	Step ID: <<The name or identifier of the start step>>

---------- END OF PAGE 1 - START OF PAGE 2 ----------

Workflow Description Specification

Workflow ID:	<<the workflow name or identifier>>	**Step ID:** <<the step name or identifier>>
Step version: <<version of this step>>	Modification date: <<date of modification>>	Step creator: <<name>>

Step

Step goal:			<<what do you want to achieve with this step?>>
Step source:			<< Is this step created locally? or it follows a reference - in that case, add link to the reference or name the person>>
Is this the first step in the workflow? Yes ☐ No ☐		Is this the final step in the workflow? Yes ☐ No ☐	
Sub-step of: <<ID of previous step (its parent)>>		Super-step of:	<<ID of next step (its child/s)>>
Order of execution:			<<e.g. first step, before Y, synchronous to Z>>
Step execution' location:			<<e.g. laboratory A, office, department, city>>
Description:			<<Action performed during this step (human action - if any)>>
Is this step concurrent/parallel to another: Yes ☐ No ☐		If yes, step ID:	<<step name or identifier>>
Standard references:			<<Standard / Approved data used for comparison e.g. Human genome >>
File Input(s):			<<Name of the necessary data to start the activity/action>>
Is the intput comming from another step: Yes ☐ No ☐		If yes, step ID:	<<step name or identifier>>
If no, what is the input's origin:			<<e.g. lab, person, tool, database>>
File Output(s):			<<Name of the generated data>>
Is the output used in another step: Yes ☐ No ☐		If yes, step ID:	<<step name or identifier>>

Tool Section

Needed tool:	<<The tool name>>
Tool version:	<<The tool's version necessary to run this step>>
Why this tool was selected:	<<Reasoning or source for the decision>>

Tool's Settings and Parameters

Loop/Repetition Section

Is this step repeated along the workflow: Yes ☐ No ☐	If yes, step ID of loop start:	<<step name or identifier>>
	If yes, step ID of loop end:	<<step name or identifier>>
If yes, how many times it repeats: <<number>>	If yes, what is needed to break the loop:	<<condition to stop the repetition>>

Condition/Threshold Section

Condition for judgment:			
Possible outcomes:	<<possibility 1 (e.g. pass, fail)>>	<<possibility 2 (e.g. pass, fail)>>	<<possibility 3 (e.g. pass, fail)>>
Next step ID:	<<the next step name for this outcome>>	<<the next step name for this outcome>>	<<the next step name for this outcome>>
Condition result:	<<e.g. send email, end flow, store data>>	<<e.g. send email, end flow, store data>>	<<e.g. send email, end flow, store data>>
Hard or soft condition:	<<Hard (a condition that was stablished and must be followed) or Soft (a condition that is good to achieve, but can be ignored)>>		

Database Section

Is the generated output stored: Yes ☐ No ☐	If yes, the data must be stored until:	<<date>>
If yes, name of the database:		<<bucket name, table name, folder name>>

GÖTEBORGS UNIVERSITET

Conclusion

- **Subjective** and **not standardised** diagrammatic & written documentation
- **First attempt** to standardise workflow documentation
- **Understandable** and **straightforward** concrete syntax extension
- **WDST** needs to be refined and automated
- **Knowledge sharing** and **formal documentation**

GÖTEBORGS UNIVERSITET

Future Work

- **Modelling tool**
  - that allows generating documentation from the diagram
  - higher precision when positioning the shapes
  - possibility to input the tool settings and parameters in the shapes
- **Validation of the concepts** with a broader bioinformatics community
- **Improvement** reduce the overloaded *control flow* shape
- **Measure**
  - if the usage of these artefacts would improve shareability and understandability
  - how many problems can be identified in the bioinformatics workflows
  - the number of manual operations that were thought automated

Questions

GÖTEBORGS UNIVERSITET

References

[1] Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2018). A brief history of bioinformatics. *Briefings in Bioinformatics*, 1-16.

[2] Kanwal, S., Lonie, A., & Sinnott, R. O. (2017, November). Digital reproducibility requirements of computational genomic workflows. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1522-1529). IEEE.

[3] Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. *Briefings in bioinformatics*, *18*(3), 530-536.

[4] Krishna, R., Elisseev, V., & Antao, S. (2018, August). BaaS: Bioinformatics as a Service. In *European Conference on Parallel Processing* (pp. 601-612). Springer, Cham.

[5] Common Workflow Language. (n.d.). Retrieved March 6, 2019, from https://www.commonwl.org/

[6] Karim, M. R., Michel, A., Zappa, A., Baranov, P., Sahay, R., & Rebholz-Schuhmann, D. (2017). Improving data workflow systems with cloud services and use of open data for bioinformatics research. *Briefings in bioinformatics*, 19(5), 1035-1050.

[7] Gray, J., & Rumpe, B. (2018). UML customization versus domain-specific languages. *Software and Systems Modeling* (SoSyM), 17(3), 713-714.

[8] Horkoff, J., de Oliveira Neto, F. G., Schliep, A., & Davila, M. (2018). *Optimized Bioinformatics Workflows from Requirement Engineering of Solution Specifications*. Unpublished report.

[9] https://software.broadinstitute.org/gatk/best-practices/workflow?id=11146

[10] D'Antonio, M., De Meo, P. D. O., Paoletti, D., Elmi, B., Pallocca, M., Sanna, N., ... & Castrignanò, T. (2013). WEP: a high-performance analysis pipeline for whole-exome data. *BMC bioinformatics*, *14*(7), S11.

[11] Marcela Davila