Comparative analysis of workflow platform in support of in silico oncology

Cite as: AIP Conference Proceedings **2172**, 020005 (2019); https://doi.org/10.1063/1.5133487
Published Online: 13 November 2019

Maria Marinova, and Vladimir Lazarov

Comparative Analysis of Workflow Platform in Support of In Silico Oncology

Maria Marinova[1, a)] and Vladimir Lazarov[2, b)]

[1]*Technical University – Sofia, Bulgaria*
[2]*European Polytechnic University, Bulgaria*

[a)]mmarinova@tu-sofia.bg,
[b)]vladimir.lazarov@epubg.eu

Abstract. In bioinformatics are used high-performance computing resources and are carried out data management and analysis tasks on large scale. A workflow consists of a set of activities which are enabled by the systematic organization of resources that analysis and process information. These systems can be useful to detailed analysis and diagnoses of breast cancer. Implementations of these frameworks differ on three elements: using friendly interface, what kind is the configuration, convention or class-based design paradigm, and offering a command line or workbench interface. In this paper we overview several workflows for bioinformatics and workflow systems to preprocess cancer-related data, like tumor/normal samples from RCGA consortium. At first, we try to give a description of workflow system, after then we give a description of scientific workflows and workflows that are used in bioinformatics. We present a modern workflow system in support of in silico oncology.

INTRODUCTION

In general, a **workflow** consists of a set of activities, which are enabled by the systematic organization of resources that transform materials, provide services, or process information. It can be depicted as a sequence of operations to complete a process.

A **Workflow Management System** (**WMS**) is software that provides an infrastructure to setup, execute, and monitor scientific workflows, concerning answers of automate complex processes on larger volumes of heterogeneous data. The WMS visualizes workflows in the form of **workflow diagrams**, depicting inputs, outputs, services and data flows. And also allows saving workflows for publishing and sharing.

CLASSIFICATION OF WORKFLOWS

Workflows exist anytime data moves from one task to another and can be structured or unstructured. There are three major types of workflows – process workflow, project workflow and case workflow. The classification is shown in Figure 1.

Proceedings of the 45th International Conference on Application of Mathematics in Engineering and Economics (AMEE'19)
AIP Conf. Proc. 2172, 020005-1–020005-7; https://doi.org/10.1063/1.5133487
Published by AIP Publishing. 978-0-7354-1919-3/$30.00

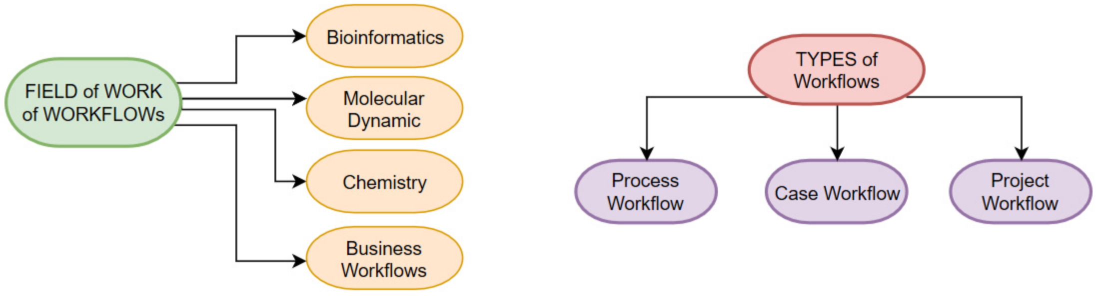


Figure 1. The criteria of classification of workflows

The process workflow is when the set of tasks is predictable and repetitive. This means that you know exactly what path it should take. Process workflow is set up to handle an unlimited number of items going through them.

In the second type of workflow - Case workflow – we don't know the path required to complete the item at the start. The path reveals itself as more data is gathered. Support tickets and insurance claims are good examples of cases. It's not clear right from the start how these items will be processed. This type of workflow can handle any number of items.

In Project workflow – we have a structured path similar to processes, but there is more flexibility here. If we want to release new version of a website, it is possible to predict with high accuracy the sequence of tasks required to complete the project. However, a project workflow is the only one good alternative.

Most resources are only referring to workflows in the sense of process workflow, but it is also important to consider the other two types.

One important property of WF is the automation. Tracking items is much easier in automated workflows. To track items in a manual workflow, you must either manually update a spreadsheet, or send lots of messages and emails to know the status. Automated workflows will show instantly where the item is in the workflow. Workflow automation has many other benefits including: eliminates redundant tasks, improves efficiency, simplifies delegation of tasks, reduces processing time, gives greater visibility, and establishes accountability. To automate your workflows, you'll need to use WF management system. Workflow management systems will allow you to create a visual representation of the workflow including all conditional tasks and exceptions.

WORKFLOW SYSTEMS FOR BIOINFORMATICS

A **bioinformatics workflow management system** is a specialized form of workflow management system designed to compose and execute a series of computational or data manipulation steps that is related to bioinformatics.

- **ANDURIL:** bioinformatics and image analysis
- **BIOBIKE**: a Web-based, programmable, integrated biological knowledge base
- **BioQUEUE**: a novel pipeline framework to accelerate bioinformatics analysis
- **Cuneiform:** A functional workflow language for large-scale data analysis
- **DiscoveryNet:** one of the earliest examples of a scientific workflow system
- **FireCloud:** a free workflow, storage, and distributed compute SaaS workspace, providing a repository of computational methods for running on the Cromwell engine.
- **GALAXY:** initially targeted at genomic
- **GenePattern:** A powerful scientific workflow system that provides access to hundreds of genomic analysis tools
- **OnlineHPC:** Online workflow designer based on Taverna.
- **Terra:** a workspace for bioinformatics, including a repository of public best practice methods, public data sets, and cloud and distributed computing facilities. Terra builds on and integrates FireCloud. Created by the Broad Institute and Verily.
- **Ugene UniPro:** provides a workflow management system that is installed on a local computer
- **VisTrails:** a workflow system very popular in the neurosciences, with many interesting workflow features

In Bioinformatics, we are faced with multiple choices of heterogeneous tools developed for an ever increasing number of applications and ever growing volumes of data. Workflow management systems (WMS) have to support physicists and biologists in the task of data processing. WMS can be classified in several ways, for example, by their paradigm (explicit or implicit), configuration system (text or scripting), workbench (command line or graphical user interface) or targeted infrastructure environment (desktop, high-performance computing, cloud). Following J. Leipzig [7] bioinformatics' workflow management systems can be estimate according some parameters. In Table 1 the common useful pipeline is shown.

TABLE 1. The modern pipeline frameworks

Syntax	Paradigm	Interaction	Example	Easy of Development	Easy of Use	Performance
Implicit	Convention	CLI	Snakemake, Nextflow, BigDataScript	★★★	★★★★★	★★★
Explicit	Convention	CLI	Ruffus, bpipe	★★★★	★★★★	★★★
Explicit	Configuration	CLI	Pegasus	★★	★★★	★★★★
Explicit	Class	CLI	Queue, Toil	★★	★★★	★★★★★
Implicit	Class	CLI	Luigi	★★★	★★★	★★★★★
Explicit	Configuration	Open Source Server Workbench	Galaxy, Taverna	★★★	★★★★	★★★
Explicit	Configuration	Commercial Cloud Workbench	DNAnexus, SevenBridges	★★	★★★★★	★★★★
Explicit	Configuratoin	Open Source Cloud API	Arvados, Agave	★★★	★★★★	★★★★

Ease of development refers to the effort required to compose workflows and also wrap new tools, such as custom or publicly available scripts and executables. Ease of use refers to the effort required to use existing pipelines to process new data, such as samples, and also the ease of sharing pipelines in a collaborative fashion. Performance refers to the efficiency of the framework in executing a pipeline, in terms of both parallelization and scalability. More stars in a column mean 'easier' or 'faster'. According тo the author J. Leipzig [8], there is no formal study of bioinformatics pipeline users, and the choice between an implicit or explicit syntax is largely a question of personal preference. Modern bioinformatics frameworks use a convention, configuration or class-based design paradigm and use an explicit or implicit syntax. Workbenches and class-based frameworks offer ease of use and performance, respectively, but require additional investment in time and expertise to integrate new tools.

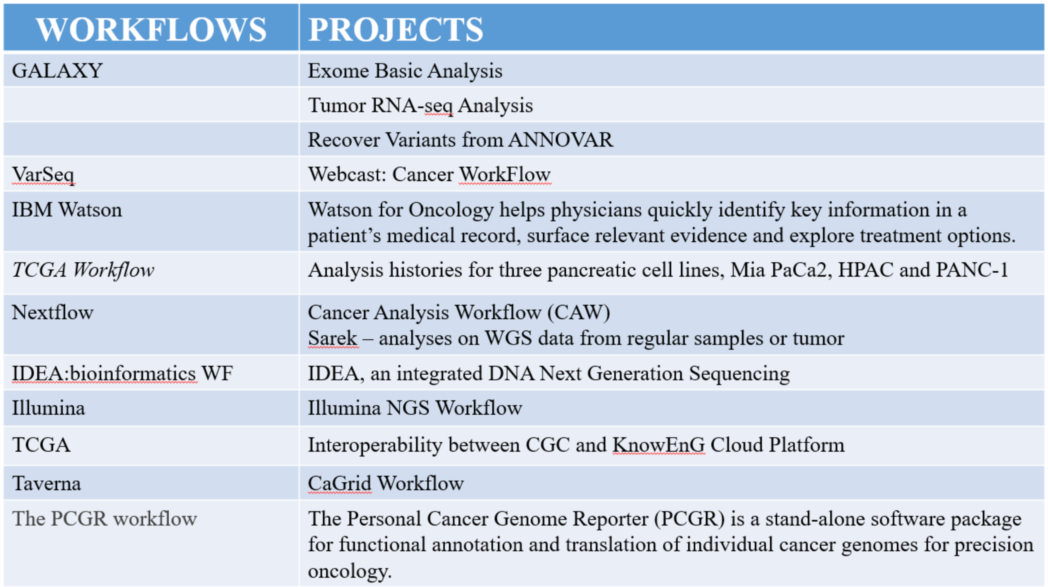
When working with BIOINFORMATICS WORKFLOWS the most important is the set of tools proposed by the WFs themselves. The following tools are necessary for testing and diagnoses of breast cancer [3,13,14]: mutation analysis, gene prediction, providing a configurable and automated framework for building VMs with biological software, next generation analysis and examination of gene expression from MicroArray data using R.

WORKFLOW SYSTEMS IN SUPPORT OF IN SILLICO ONCOLOGY

Using workflows the doctors can perform analysis of different cancer as shown in Table 2. The Cancer Genome Atlas (TCGA) is a public funded project that aims to catalogue and discover major cancer-causing genomic alterations to create a comprehensive "atlas" of cancer genomic profiles. So far, TCGA[15] researchers have analyzed large cohorts of over 30 human tumors through large-scale genome sequencing and integrated multi-dimensional analyses. A major goal of the project was to provide publicly available datasets to improve diagnostic methods, treatment standards, and finally to prevent cancer. The workflow Sarek [28] is a workflow tool designed to run analyses on WGS data from regular samples or tumor / normal pairs, including relapse samples if required. It's built using the workflow Nextflow, bioinformatics domain specific language for workflow building. Software dependencies are handled using Docker and Singularity_– container technologies that provide excellent reproducibility and ease of use.

Illumina offers a comprehensive solution for the NGS workflow, from library preparation to data analysis [21]. Library preparation kits are available for all NGS methods, including WGS, exome sequencing, targeted sequencing, RNA-Seq, and more.

TABLE 2. Workflow system in support of in silico oncology



WORKFLOWS	PROJECTS
GALAXY	Exome Basic Analysis
	Tumor RNA-seq Analysis
	Recover Variants from ANNOVAR
VarSeq	Webcast: Cancer WorkFlow
IBM Watson	Watson for Oncology helps physicians quickly identify key information in a patient's medical record, surface relevant evidence and explore treatment options.
TCGA Workflow	Analysis histories for three pancreatic cell lines, Mia PaCa2, HPAC and PANC-1
Nextflow	Cancer Analysis Workflow (CAW) Sarek – analyses on WGS data from regular samples or tumor
IDEA:bioinformatics WF	IDEA, an integrated DNA Next Generation Sequencing
Illumina	Illumina NGS Workflow
TCGA	Interoperability between CGC and KnowEnG Cloud Platform
Taverna	CaGrid Workflow
The PCGR workflow	The Personal Cancer Genome Reporter (PCGR) is a stand-alone software package for functional annotation and translation of individual cancer genomes for precision oncology.

By extending the Taverna Workbench, CaGrid Workflow Toolkit provided a comprehensive solution to compose and coordinate services in caGrid, which would otherwise remain isolated and disconnected from each other. Using it users can access more than 140 services and are offered with a rich set of features including discovery of data and analytical services, query and transfer of data, security protections for service invocations, state management in service interactions, and sharing of workflows, experiences and best practices. The one of used application for this WF is cancer (caBIG)[22] .

On the next Table 3, some useful workflows for in silico oncology are shown. The first workflow is Galaxy which is a scientific workflow system. These systems provide a means to build multi-step computational analyses. They typically provide a GUI and are a platform for biological data. It supports data uploads from the user's computer, by URL, and directly from many online resources (such as the UCSC Genome Browser, BioMart and InterMine). Galaxy supports a range of widely used biological data format, and translation between those formats. Galaxy provides a web interface to many text manipulation utilities, enabling researchers to do their own custom reformatting and manipulation. Many biological file formats include genomic interval data (a frame of reference, e.g., chromosome name, and start and stop positions), allowing these data to be integrated. Another advantage of Galaxy is open-source software, implemented using Python.

Taverna [2,17] is an open source and domain-independent Workflow Management System – a suite of tools used to design and execute *scientific workflows* and aid *in silico* experimentation. Taverna Workbench Bioinformatics is an edition of Taverna Workbench that includes support for building and executing bioinformatics workflows using bioinformatics data and analytical services such as BioMart and BioMoby. Taverna Workbench Bioinformatics is equivalent to the Taverna Workbench Core bundled with BioMart, BioMoby, SoapLab and API Consumer plugins. Taverna Workbench Bioinformatics is distributed under the open source license LGPL 2.1.

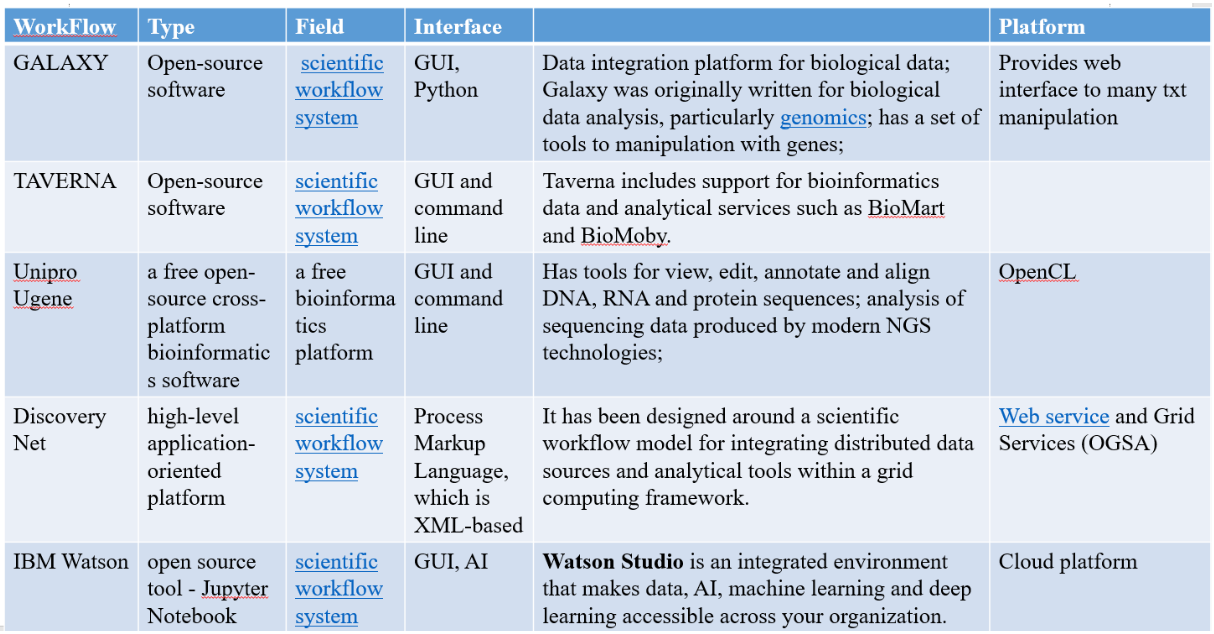
Unipro UGENE is a free open-source cross-platform bioinformatics software. It allows you to view, edit, annotate and align DNA, RNA and protein sequences, work with 3D structures and surface algorithms and model workflows using the **Workflow Designer**. Overview of NGS Pipelines in UGENE is a free bioinformatics platform that integrates dozens of well-known biological tools and algorithms and provides both graphical and command line interfaces for them. It is designed to solve various complex computational tasks in bioinformatics. One of the areas is analysis of sequencing data produced by popular modern NGS technologies.

UGENE makes NGS data analysis easier. There are many pipelines in this workflow. Three popular pipelines for analyzing NGS data extend the UGENE NGS framework: Variant calling with SAMtools, RNA-Seq data analysis with Tuxedo, ChIP-seq data analysis with Cistrome.

The Ugene Unipro workflow supports OpenCL and you can use it to speed up some calculations. New and powerful next-generation sequencing (NGS) techniques allow to simultaneously and quickly analyze a large number of genes, up to the entire genome , that are assumed to be involved in diseases. The main challenge in applying NGS to medical diagnosis resides in workflow development fulfilling diagnosis interpretation requirements, such as quality control or variant knowledge annotation.

Discovery Net [10,19] is one of the earliest examples of a scientific workflow system allowing users to coordinate the execution of remote services based on Web service and Grid Services standards. The workflow management system has been designed around a scientific workflow model for integrating distributed data sources and analytical tools within a grid computing framework. During 2001-2005 Discovery Net[19] was developed as a part of UK-e-Science project and had aim to produce a high-level application-oriented platform, focused on the user scientists in deriving new knowledge from devices, sensors, databases , analysis components and computational resources that reside across the Internet or grid. Workflows in this system are represented and stored, using Discovery Process Markup Language, which is XML-based representation language for workflow graphs. The type of date by default in the system is a relational table. However for supporting bioinformatics data model was added.

TABLE 3. Scientific Workflows - parameters and characteristics.



WorkFlow	Type	Field	Interface		Platform
GALAXY	Open-source software	scientific workflow system	GUI, Python	Data integration platform for biological data; Galaxy was originally written for biological data analysis, particularly genomics; has a set of tools to manipulation with genes;	Provides web interface to many txt manipulation
TAVERNA	Open-source software	scientific workflow system	GUI and command line	Taverna includes support for bioinformatics data and analytical services such as BioMart and BioMoby.	
Unipro Ugene	a free open-source cross-platform bioinformatics software	a free bioinformatics platform	GUI and command line	Has tools for view, edit, annotate and align DNA, RNA and protein sequences; analysis of sequencing data produced by modern NGS technologies;	OpenCL
Discovery Net	high-level application-oriented platform	scientific workflow system	Process Markup Language, which is XML-based	It has been designed around a scientific workflow model for integrating distributed data sources and analytical tools within a grid computing framework.	Web service and Grid Services (OGSA)
IBM Watson	open source tool - Jupyter Notebook	scientific workflow system	GUI, AI	**Watson Studio** is an integrated environment that makes data, AI, machine learning and deep learning accessible across your organization.	Cloud platform

The most important characteristic when choosing appropriate workflow system remains the realization of WMS, which imposes big influence on performance. According the researchers Elise Larsonneur and Jonathan Mercier [5], the implementation of workflow management systems can significantly affect their overall performance.

The choice of the workflow management system and related environment will condition the ease of development, as well as compatibility and sharing. However, it remains difficult to obtain a comprehensive view of the computing efficiency of this workflow management system. The detail analysis and research is made in paper “Evaluating Workflow Management Systems: A Bioinformatics Use Case”. The authors evaluate some popular WMS and workflow languages — Cromwell-WDL, Nextflow, Pegasus-mpi-cluster, Snakemake , Toil-CWL. They pretend in their paper that Snakemake showed average usage of CPU and was the fastest workflow management system to process the workflow. Snakemake implementation (python-based as Toil-CWL) also showed frequent voluntary context switches during its execution, although to a lesser extent.

CONCLUSION

As a result of our analyses we propose some consequences of steps to realize a workflow concerning the breast cancer diagnosis and therapy. Each step is realized by one or several pipelines, created from mixed contribution of original applications and workflow management system. For example, one possible picture of some common contribution to assist breast cancer diagnosis and therapy might be as follows:

- First step – diagnosis, consisting of two pipelines: mammography and biopsies. The first one is defining the presence of illness, and eventually invoke the second one;
- Second step – predictive, consisting again of two pipelines: DNA analysis to discover possible gene mutation and risk, and therapy, invoked by the first one pipeline /we suppose that the therapy is already necessary/. The second pipeline is very powerful one, allowing decision making for successful treatment of patients.

AKNOWLEDGMENTS

This paper presents the outcomes of research project "Intelligent Method for Adaptive In-silico Knowledge Discovery and Decision Making Based on Analysis of Big Data Streams for Scientific Research", contract ДН07/24, financed by the National Science Fund, Competition for Financial Support for Fundamental Research – 2016, Ministry of Education and Science, Bulgaria.

REFERENCES

1. A. Von Eschenbach, K. Buetow, Cancer Informatics Vision: caBIG". Cancer Informatics 2006.
2. D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, T. Oinn, "Taverna: a tool for building and running workflows of services", 2006 Nucleic acids research, 34:W729-W732.
3. D. Ivanova, P. Borovska, S. Zahov, Development of PaaS using AWS and Terraform for medical imaging analytics, AIP Conference Proceedings, Volume 2048, 10 December 2018, Article number 060018, DOI: 10.1063/1.5082133
4. E. Deelman, "Pegasus: A framework for mapping complex scientific workflows onto distributed systems". Scientific Programming 2005, 13:219-237.
5. E. Larsonneur, J. Mercier, et al., "Evaluatinig Workflow Management Systems: A Bioinformatics Use Case", IEEE International Conference on Bioinformatics and Biomedicine 2018.
6. G. Corti, A. Bartolini, G. Crisafulli, etc. "A genomic analysis workflow for colorectal cancer precision oncology."
7. J. Elhai, A. Taton, J. Massar, J. K. Myers, M. Travers, J. Casey, M. Slupesky, J. Shrager, "BioBIKE: A Web-based, programmable, integrated biological knowledge base", 2009, Nucleic Acids Research. 37 (Web Server issue): W28–W32.
8. J. Leipzig, "A review of bioinformatics pipeline frameworks", Briefings in Bioinformatics, vol, 18, issue 3, 2017, pp530-536.
9. J. Saltz, S. Oster, S. Hastings, S. Langella, T. Kurc, W. Sanchez, M. Kher, A. Manisundaram, K. Shanbhag, P. Covitz, "caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid". Bioinformatics 2006, 22:1910-1916.
10. M. Ghanem, et al., "Building and using analytical workflows in Discovery Net." N Data mining on Grid (ed. W. Dubitzky), 2008, pp. 119-140.
11. N. Fortier, "VARSEQ: Cancer Gene Panels and Tumor-Normal Workflows", 2016, www.blog.goldenhelix.com
12. P. Di Tommaso st al., "Nextflow enables reproducible computational workflows," Nature Biotechnology, vol. 35, no. 4, pp. 316–319, 2017.
13. P. Borovska, "Big Data Analytics and Internet of medical Things Make Precision Medicine a Reality", International Journal of Internet of Things and Web Services, Volume 3, 2018, pp. 24-31, ISSN: 2367-9115, http://www.iaras.org/iaras/journals/ijitws
14. P. Borovska, V. Gancheva and I. Georgiev, "Platform for adaptive knowledge discovery and decision making based on big genomics data analytics," In: Rojas I., Valenzuela O., Rojas F., Ortuño F. (eds) Bioinformatics and

Biomedical Engineering. IWBBIO 2019, *Lecture Notes in Computer Science*, vol. 11466. Springer, Cham, pp. 297-308, https://doi.org/10.1007/978-3-030-17935-9_27.

15. P. Romano, E. Bartocci, et al. “Biowep: a workflow enactment portal for bioinformatics applications”, BMV Bioinformatics 2007,8(Suppl I):S19
16. S. Angiuoli, M. Maralka, A. Gussman, K. Galens, M. Vangala, Dr. Riley, C. Arze, J. White, O. White, “CloVR WFFricke: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing”. BMC Bioinformatics 2011, 12:356.
17. T. Oinn, M. Greenwood, M. Addis, M. N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. R. Pocock, M. Senger, R. Stevens, A. Wipat, C. Wroe, "Taverna: Lessons in creating a workflow environment for the life sciences" 2006.
18. T. Silva, et al., “TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages”, F1000Research 2016, 5:1542.
19. V. Curcin, M. Ghanem, ” Scientific workflow systems - can one size fit all?”, 2008, Cairo International Biomedical Engineering Conference. pp. 1–9.
20. V. Curcin, M. Ghanem, Y. Guo, “The design and implementation of a workflow analysis tool”, Phil. Trans. R. Soc A(2010) 368, 4193-4208.
21. https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf
22. myExperiment – caBIG workflows. [http://www.myexperiment.org/search?query=cabig&type=workflows].
23. http://www.taverna.org.uk/download/workbench/2-5/bioinformatics/
24. https://usegalaxy.org/u/jeremy/p/cancer-analyses
25. https://github.com/genome/analysis-workflows
26. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5302158/
27. https://www.ncbi.nlm.nih.gov/pubmed/25691825
28. https://opensource.scilifelab.se/projects/sarek/
29. https://ugene.net