BIOCYBERNETICS AND BIOMEDICAL ENGINEERING 41 (2021) 1646–1663

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bbe

Original Research Article

Integration of solutions and services for multi-omics data analysis towards personalized medicine

Daniel Reska [a], *Marcin Czajkowski* [a,*], *Krzysztof Jurczuk* [a], *Cezary Boldak* [a], *Wojciech Kwedlo* [a], *Witold Bauer* [b], *Jolanta Koszelew* [a], *Marek Kretowski* [a]

[a] *Faculty of Computer Science, Bialystok University of Technology, Wiejska 45A, 15-351 Bialystok, Poland*
[b] *Clinical Research Centre, Medical University of Bialystok, Sklodowska – Curie 24A, 15-276 Bialystok, Poland*

ARTICLE INFO

Article history:
Received 19 March 2021
Received in revised form
6 October 2021
Accepted 12 October 2021
Available online 2 November 2021

Keywords:
Decision support
Precision medicine
Multi-omics
Big data

ABSTRACT

Current advances in high-throughput and imaging technologies are paving the way next-generation healthcare, tailored to the clinical and molecular characteristics of each patient. The Big Data obtained from these technologies are of little value to society unless it can be analyzed, interpreted, and applied in a relatively customized and inexpensive way. We propose a flexible decision support system called IntelliOmics for multi-omics data analysis constituted with well-designed and maintained components with open license for both personal and commercial use. Our proposition aims to serve some insight how to build your own local end-to-end service towards personalized medicine: from raw data upload, intelligent integration and exploration to detailed analysis accompanying clinical medical reports.

The high-throughput data is effectively collected and processed in a parallel and distributed manner using the Hadoop framework and user-defined scripts. Heterogeneous data transformation performed mainly on the Apache Hive is then integrated into a so called 'knowledge base'. On its basis, manual analysis in the form of hierarchical rules can be performed as well as automatic data analysis with Apache Spark and machine learning library MLlib. Finally, diagnostic and prognostic tools, charts, tables, statistical tests and print-ready clinical reports for an individual or group of patients are provided. The experimental evaluation was performed as part of the clinical decision support for targeted therapy in non-small cell lung cancer. The system managed to successfully process over a hundred of multi-omic patient data and offers various functionalities for different types of users: researchers, bio-statisticians/bioinformaticians, clinicians and medical board.

* Corresponding author.
E-mail address: m.czajkowski@pb.edu.pl (M. Czajkowski).

https://doi.org/10.1016/j.bbe.2021.10.005

1. Introduction

Life science is becoming increasingly collaborative and complex. Scientists use continuously emerging and diverse technologies to better understand organisms and diseases from a molecular to the system level. Researchers need to explore heterogeneous data obtained from multiple sources to conduct meaningful analyses and extract actionable knowledge. In recent years, an increase in data size and availability has been particularly dramatic due to the 'omics' technologies. The term Big Data [1] is often used in such cases, which refers to high volume, high velocity, and/or high variety information assets. In the context of bio-medicine, we also face various data representations, high-dimensionality, and incompleteness or imprecision of observations. Recent advances in the interdisciplinary fields of precision medicine, data mining and predictive algorithms, bioinformatics, and computational medicine are remarkable [2,3]. Nonetheless, there is a widening gap between data acquisition (due to the rapid technological progress) and the comparatively slow functional characterization of biomedical information. In this regard, new ideas of improving the general access to efficient molecular information management, integration, analysis, and interpretation are becoming crucial.

One of the important applications of precision medicine is clinical oncology, as the exact mechanisms of carcinogenesis in its prognosis are often unclear. Multi-omics technologies are being widely used to systematically understand the formation of cancer on different biological levels by focusing on multi-parameters systematical models [4]. Cancer is a complex, whole-body disease and involves multiple abnormalities in the levels of DNA, RNA, and other molecules including proteins. The most commonly used data types that compose multi-omics models [5] are:

- genomics – to identify the nucleotide variants (SNPs – single nucleotide polymorphisms) in the whole genome associated with clinical traits (GWAS – genome-wide association study); technology/platform: genotyping arrays and whole-exome sequencing;
- transcriptomics – to quantify the expression levels of cellular transcripts (e.g. mRNA); technology/platform: expression arrays, RNA sequencing;
- proteomics - to characterize the protein expression levels of cells/samples; technology/platform: mass spectrometry (MS) -based approaches;
- metabolomics – to characterize the abundance profile of metabolites and their relative ratios; technology/platform: alike in proteomics;
- radioomics - to quantify the features of medical imaging; technology/platform: CT, MRI, PET;
- others like epigenomics, microbiomics, exposomics, etc.

Multi-omics data and technologies have a wide-range of applications in both basic research and clinical treatment of cancer [6]. The term "cancer genomics" [7] refers to the study of tumor genome. Its goal is to survey multi-omics data to identify genes and pathways deregulated in cancer, and reveal those that may be used as the early stage biomarkers of the disease and help understand the pathogenesis of cancer. Such discoveries improve our understanding of the biology of cancer and may lead to the discovery of novel diagnostic, prognostic, and therapeutic biomarkers that will ultimately improve disease detection and treatment. Cancer genomics are rapidly evolving and coupled with the ever-increasing efficiency of genomic profiling. This leads to the fact that personalized medicine is likely to become the reality soon. It is expected that in the near future oncological patients will be profiled in a timely manner, and that the multi-omics findings will subsequently be introduced into clinical practices.

Several bottlenecks slow-down the transition from conventional to personalized medicine [8]. Efforts to integrate heterogeneous big data, including data coming from the multi-omics technologies (such as molecular and protein signatures, individual genome sequences, patients' clinical phenotypes, images and follow-up data) is a tough task [9]. Most omics data are only qualitative in nature, making it very hard to reproduce and even harder to compare. The lack of sufficient meta-data is also a roadblock to the successful integration of multi-omics data sets. Often very little effort goes into collecting the meta-data about the samples or patients. To enable reproducibility and biologically relevant interpretation of omics results the information about the observable phenotypes of the samples should also be collected.

Currently existing digitized data and clinical information are present in multiple formats and are largely unstructured. In the absence of a universally accepted standards, institutions continue to generate silos of information [10] in various formats and store them in heterogeneous databases. Similarly, it is common for bioinformatic programs to require input data in non-standardized formats and output results data in a format that is incompatible with other programs. This can make it difficult to create a multi-program or a multi-database workflow and may require users to spend their time writing scripts to convert and reconvert data so that it can be read by other programs.

Many of the specific analytical tools and experimental designs traditionally used for individual omics disciplines (e.g., genomics, transcriptomics, and proteomics) are not sufficiently well-suited to permit proper comparisons or intelligent integration [11]. Often, other types of data like advanced clinical or imaging are not exploited or embedded in the systems. Nonetheless, there might exist many well designed and maintained software tools that are often unaware of by the researchers. This problem may be due to the sheer number of resources or the lack of a central repository which catalogs, links and rates or summarizes these tools according to e.g. their functionalities, limitations and difficulty of using. Treating every problem with the same tool may be partly caused by the low user-friendliness of many currently available multi-omics programs.

Finally, multi-omics commercial products are expensive and require a considerable level of funding for the software maintenance. Therefore, local research projects and medical centers seek for a software with the free license to cut the costs. The vast majority of such solutions offer narrowly targeted products that are addressed to research, rather than to hospital centers, thus no real medical decision support can be proposed [11]. There is a strong demand for systems designed and implemented in close cooperation with medical centres. The most of currently available systems and tools are not accessible by physicians or researchers without significant training in data science [4]. Usability and specificity of the solutions for a particular problem are often overlooked, which strongly affects the everyday use of the existing systems in the clinical practices. As a result, very often an invaluable knowledge and new discoveries coming from the integrated analyses are difficult to access [12]. It seems simply infeasible to gather and make a proper comparison or review all existing solutions. Therefore, in Tables 1 and 2, we only show aggregated information from recent reviews concerning software and web tools, databases, AI algorithms, projects in the context of precision medicine.

Inspiration for this research is a Databricks solution [41] - one of the largest platforms for massive-scale data engineering, machine learning and business analytics. Databricks is build on the Apache Spark [42] which is the largest open-source cluster computing framework and unified analytics engine for big data processing. Multiple other open-source components like DataLake, MLFlow and TensorFlow are connected with the Databricks which wrap them up into an unified data web-based platform. The main downside is that the solution is relatively expensive and lacks of flexibility in contrast to open-source components that constitute it.

Our goal was somewhat similar to the founders of Databricks as we wanted to design a web-based decision support system, composed of available top open-source solutions, for multi-omics data analysis. Within a few years we have developed the intelligent solutions and services for multi-omics data system (IntelliOmics) tailored for precision medicine within the medium-scale medical project [43]. The platform offers a complete and flexible approach starting from the raw data upload to the patient (or group of patients) medical reports, designed according to the specific clinicians' needs. It is based on a client–server architecture, with computational and data storage modules deployed on the server, and browser-based client applications [44]. In this paper we present our way to build such a system and share guidelines for all those who are considering creating a solution similar to IntelliOmics.

2. Material and methods

In this paper, we present a self-contained web-based system called IntelliOmics that allows a complete analysis starting from the raw data (text, images, binary files), all the way to the diagnostic reports which can be associated with the treatment recommendation tailored to the clinicians needs. Fig. 1 illustrates three main components in which the system is organized: *data management*, *data processing*, and *data analysis*. Table 3 shows the main user groups, available functionalities and interfaces within the system.

2.1. *Data management*

The first component required by any personalized medicine solution is the upload and the storage of clinical and multi-omics data from different sources. The typical system should enable repository of the original, unprocessed data, as well as intermediate, processed data, and the results of the analyses.

It should be emphasized that especially various omics data demands massive amounts of computing power and space [52]. A typical single next-generation sequencing method can generate up to 300 GB of data. Even after data preprocessing, including e.g. genome mapping, there is approximately 2–3 GB of the heterogeneous data (including report files, intermediate results, potential biomarkers list, genomic structural and functional variants saved into vcf files). Other types of data related to the transcriptomics, proteomics, metabolites and e.g. PET/MRI imaging are accordingly smaller but still large for storage and processing. Therefore, structures and technologies for both Big Data and classical data warehousing should be applied.

As the web frameworks are crucial to app-development [53], one needs to be very attentive and serious about selecting the right development framework. We have identified a long list of requirements and constraints that must be met and used them to build the IntelliOmics system. We have selected Django [54], which is considered one of the best high-level Python web framework. It is not only free and open source, but also exceedingly scalable and reassuringly secure. It encourages rapid development and is especially useful for creating database-driven websites. It comes with an advanced user authentication system that also handles specific permissions for users, groups, data etc. This way we could easily set IntelliOmics users (or groups of users) permissions to a part of the stored data, functionalities or interface that they need.

IntelliOmics environment aims at safe and reliable collection, storage and intelligent processing of clinical and multi-omics data of patients qualified to the considered groups. The groups can be defined based on e.g., diagnosis. Staging and the group assignment can be latter modified. Currently, IntelliOmics accepts the data from:

- personal and clinical information (unstructured survey data of dietary and lifestyle habits, clinical medical records, medical examination results including e.g. complete blood count, hormones, drug use and body composition);
- results from analysis of biological materials including blood, other liquid and tissue samples;
- PET/MRI and CT images (raw & DICOM images with descriptions);
- whole genome sequencing (plus DNA methylation) data in raw or already transformed formats;

Table 1 – Aggregated information from recent literature reviews on software and data in context of precision medicine.

Category	Reference	Content and comparisons	Number of elements
Software tools	Pinu, F.R., et al. [13]	integrated omics, domain, functionality, type of licence	**40** multi-omics integration software tools and web applications
	de Anda-Jáuregui, et al. [14]	details, features and notes, use cases	**86** tools for multi-omics computational oncology
	Huang, S., et al. [15]	category, data type, output, methods	**32** supervised and unsupervised methods for the multi-omics data integration
	Github pages [16]	open-source list of existing workflow systems	**285** computational data analysis workflow systems
	Chung, R., et al. [17]	OmicsSIMLA tool for simulating multi-omics data	**4** 4 types of omics data used in simulation
	Zanfardino, M., et al. [18]	GUI for integrated multi-omics data in radiogenomic studies	TCGA-BRCA data collection
Web solutions	Misra, B.B., et al. [19]	platform, degree of user friendliness, functionalities and availability	**65** tools & software for integrated –omics analysis
	Subramanian, I., et al. [20]	omics data supported, source repository, availability for private data	**9** multi-omics data analysis and visualization portals
Databases	Pinu, F.R., et al. [13]	integrated omics, domain, functionality, type of license	**15** databases that aid multi-omics data integration process
	Subramanian, I., et al. [20]	diseases and type of multi-omics data available	**17** multi-omics data repositories
	Labory, J., et al. [21]	general and MD (mitochondrial) omics databases	**17** MD databases, **21** general databases
	Rappoport N., et al. [22]	the Cancer Genome Atlas (TCGA) datasets used in the clustering methods	**9** methods bencharked on **10** TCGA datasets
	Lee B., et al. [23]	novel computational methods for the inference of novel biological relations from the heterogeneous multi-layered network (HMLN)	4 groups of algorithms: matrix factorization, random walk, meta-path, graph convolutional network (GCN)
Artificial intelligence approaches	Zeesjam, A., et al. [24]	approaches, objectives and AI&ML reviews	**32** feature and variability analysis of reviewed approaches, **11** real time examples of AI and ML algorithms
	Cirillo, D., et al. [25]	tasks, data type, methods	**9** notable biomedical applications of deep learning
	Tong, L., et al. [26]	survival risk analysis	**4** TCGA cancer datasets from UCSC Xena, each with four -omics data
	Gambardella, V., et al. [27]	limitions of molecular driven threatmens in the clinic as well as methods to overcome these limitations	3 ways to overcome the limitations of current genomics approach, 7 clinical trials and corespoding molecular tools
	Su, M., et al. [28]	recent technological achievements in proteomics, the key hallmarks of cancer, and unmet clinical needs	5 proteomics toolboxes, 10 hallmarks soft cancer, 10 cancer biomarkers discovered by proteomics
Projects	Patil, S., et al. [51]	biobanks, localization, collaborations	**9** worldwide biobanks and their collaborative projects
	Cirillo, D., et al. [25]	initiatives and research focuses	**9** large international consortia and **23** ongoing population-scale sequencing initiatives for Personalized Medicine

Table 2 – Aggregated information from recent medical and biological literature in context of precision medicine.

Reference	Content and comparisons	Experiments
Morello, G., et al. [29]	demonstrating genomic, transcriptomic, proteomic and other factors linked to ALS, arguing for necessity of holistic approaches, integrating mnultiple data types of omics data	37 genes linked to ALS, 32 proteins
Hou, X., et al. [30]	only describing genomic (monogenic and poligenic-GWAS), transcriptomic and epigenomic factors related to JIA, but nothing on integrating these data	27 SNPs related to JIA from GWAS analysis, 4 genes correlated with monogenic form of JIA
Song J., et al. [31]	analysis plasma lipidome and metabolome allows to distinguish COVID-19 patients from healthy controls	–
Miki, D., et al. [32]	GWAS for virus-related hepatocellular carcinoma (HCC), outline for future HCC personalized treatments	one associated SNP found in the GWAS
Rivenbark, A.G., et al. [33]	review of molecular and cellular heterogeneity of breast cancer in context of classification and personalized treatment	5 molecular assays for breast cancer assesment mentioned
Krzyszczyk, P., et al. [34]	challenges in precision and personalized medicine in cancer treatment: general issues in cancer treatment, acquiring and storage of omics, physiological and lifestyle data, therapy development, regulational and ethical issues	–
Frohlich, H., et al. [35]	review of state-of-the-art data science approaches for personalized medicine, challenges, and future directions	–
Couri, T., et al. [36]	Hepatocellular carcinoma therapies	9 molecular targets for HCC therapies
Zanfardino, M., et al. [37]	Integration of radiomics into multi-omics framework. R package MultiAssayExperiment used to store multi-omics data	36 radiomic features extracted from primary tumor images of 91 patients
Zeng, Y., et al. [38]	Identification of potential biomarkers and therapeutical targets associated with retinoblastoma	193 differentially expressed genes, 74 differentially methylated-differentially expressed genes, 45 differentially expressed miRNAs and 5 differentially expressed IncRNAs identified
Lawal B., et al. [39]	cancer-associated fibroblast and immune infiltrations based on gene expressions and alterations	total of 56 cohorts (tumor vs normal) compared
Xie, F., et al. [40]	bio-printing of primary human hepatocellular carcinoma for personalized medicine	HCC models in vitro for 6 patients

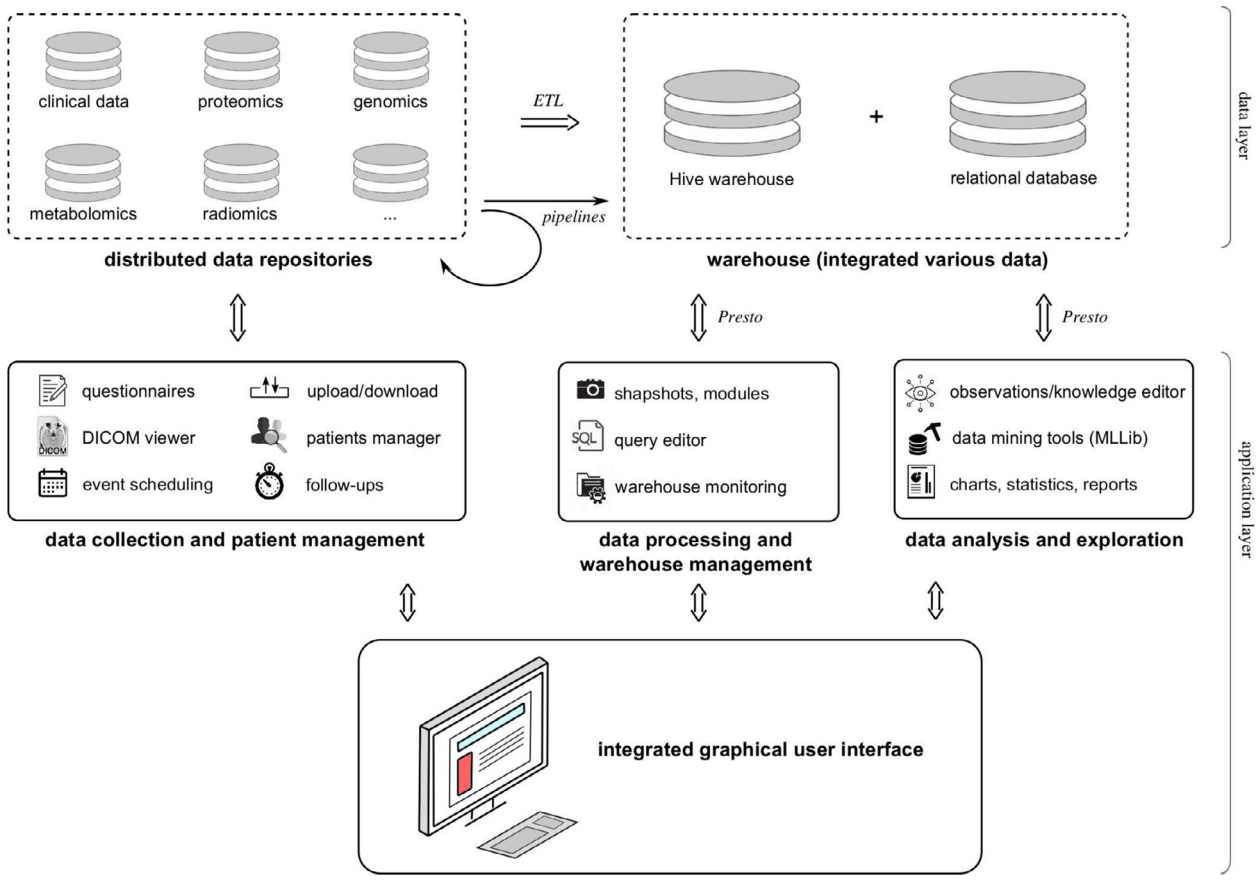

Fig. 1 – IntelliOmics system architecture.

- transcriptomic data, smallRNA fraction in raw or already transformed formats including e.g. expression levels of specific genes;
- metabolomic and proteomic data that are generated by mass spectrometers.

Structures and technologies for both Big Data and classical data warehousing are applied. We decided to use a relational PostgreSQL database to store all clinical data and the localization in the file-system of the distributed omics data repositories. The data load module may be run at the web application level, but a direct data transfer is also possible, especially for WGS data.

In order to ensure the flexibility of the proposed solution we have developed a mechanism that enables the users to create survey forms with various types of data describing the patient, disease and medical examination. Within the data upload, it is possible to define the specific characteristics of the data format. Because the IntelliOmics system is designed to store very sensitive data, a particular attention is focused on its confidentiality and security. Each data chunk is linked to the unique anonymized patient's code and time stamp. The data can also be associated with additional fully-customized questionnaires. They can be configured within the interface by using additional free Django packages.

In the data management module, it is possible to define a new workflow (set of rules, including calendar events planning and examination appointment) for an individual and a group of patients. It is also possible to use the predefined workflows. Medical and clinical staff can plan patient's follow up visits and procedures using a calendar module. Additionally, system provides up-to-date detailed reports on the acquired data and planned medical examinations and other functionalities described in Table 3. Example screenshots of the IntelliOmics data management interface are illustrated in Fig. 2.

2.2. *Data processing*

Reliable storage and efficient processing of hundreds of terabytes of data is not a trivial task [55]. Usually, storage and processing of huge amount of data demands the use of parallel and distributed tools. Currently, the most popular open source technology that allows creating low-cost and high-performance solutions for Big Data is the Apache Hadoop [56] ecosystem. It refers to numerous components of the Apache Hadoop software library and includes open source projects as well as a complete range of complementary tools. Some of the most well-known tools of Hadoop ecosystem include HDFS, Hive, YARN, MapReduce, Spark, etc. Here are

Table 3 – IntelliOmics users, functionalities and interfaces.

Component	User group	Interfaces	Functionalities	Difficulty
Data management	physicians, medical assistants	questionnaires	filling individual patient's survey data, questionnaires, consents etc.	low
	physicians, medical assistants	data storage	single patient's upload through web UI (various data-specific interfaces)	low
	physicians	calendar	events planning and medical examination appointment (patient, location, study perspective)	low
	physicians	management	view patients data	low
	medical staff, scientists	data storage	multiple patients upload through web UI and a direct data transfer (batch)	medium
	medical staff	events	define event types (upload specific file, fill a form, contact etc.) that can be applied for a patient	medium
	medical staff	templates	define a new workflow (set of rules, including calendar events planning and examination appointment) for an individual and a group of patients	medium
	medical staff	reports	view up-to-date detailed reports on the acquired data and planned medical examinations	medium
Data processing	administrator	warehouse monitoring	manage and monitor Apache Hadoop clusters (hardware and software)	medium
	administrator	snapshots	designing and running snapshots of selected data type and scheme	high
	bioinformaticians	workflows	designing pipelines with external predefined scripts and programs	high
	bioinformaticians, scientists	transformations	designing and running data transformations pivoting, normalization, cleaning, simple integration etc.	high
	scientists	workflows	view and run data workflows, check status of data preprocessing	medium
Data analysis	bioinformaticians, scientists	query editor	view and test queries using SQL-like editor	high
	bioinformaticians, scientists	knowledge editor	add and manage expert knowledge in form of rules (observations)	medium
	bioinformaticians, scientists	data mining	manage and perform data analysis based on raw or transformed data	high
	scientists, physicians	statistics, charts, tables	view or export information gathered in the knowledge base or returned by data mining module	low
	physicians	DICOM viewer	view and mark patients imaging data through DICOM viewer	low
	physicians	reports	extract patients information and diagnostic results of analysis in form of a clinical reports (in pdf)	low

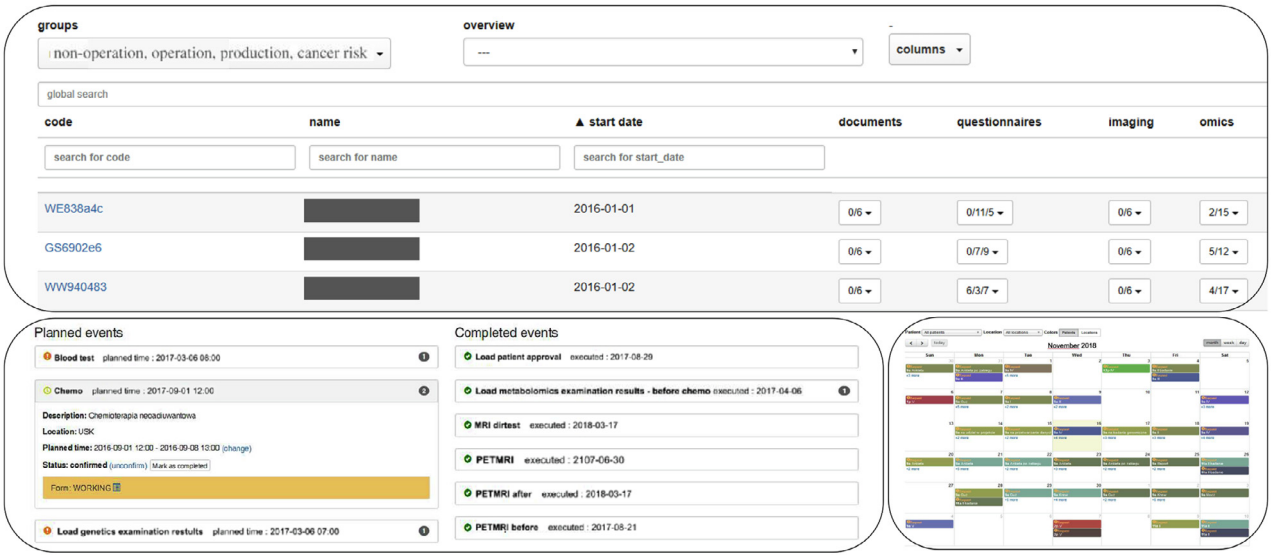

Fig. 2 – User interfaces of the data management component.

the elements that compose IntelliOmics and also are frequently used by the developers:

- HDFS (Hadoop Distributed File System) is the primary storage system of Hadoop. It is a distributed file system able to store large files running over the cluster of commodity hardware. It employs a master (NameNode) and workers (DataNodes) architecture;
- MapReduce is the main data processing layer of Hadoop. It has the capability to process large structured and unstructured data as well as to manage very large data files in parallel by dividing the job into a set of independent tasks (sub-jobs);
- YARN (Yet Another Resource Navigator) is one of the core components suitable for resource management. It is responsible for managing workloads, monitoring, and security controls implementation. It also allocates system resources to the various applications running in a Hadoop cluster while assigning which tasks should be executed by each cluster nodes;
- Hive [57] is an Extract, Transform, Load (ETL) and Data warehousing tool used to query or analyze stored datasets. Hive has three main functions: data summarization, query, and analysis of unstructured and semi-structured data in Hadoop. It features a SQL-like language (HQL) that automatically translates queries into distributed jobs based on MapReduce or other engines;
- Spark [42] is a distributed computing engine for in-memory Big Data processing. Spark can be deployed in several ways, it supports Java, Python, Scala, and R programming languages, and provides SQL analysis, data streaming, machine learning, and graph processing modules, which can be used together.

In context of Big Data processing, we follow the top commercial solutions, including, Databricks, which are built on top of Hadoop ecosystem. The second component of the IntelliOmics (see Fig. 1), moves uploaded data into Hive as the data warehouse software. The Hive efficiently stores the data and performs large-scale ETL operations that excees the capabilities of traditional relational warehouses. Additionally, some ETL work can be also performed by Spark [42].

Finally, the multi-omics data requires advanced processing, mainly the data cleaning, QC, and the data transformation (normalization, pivoting, etc.). The system should be open to data-specific transformations that require many calculation-intensive steps with the use of various external, constantly changing tools. There are hundreds of free computational data analysis workflow systems [16]. Currently, one of the most popular bioinformatics workflow manager that enables the development of portable and reproducible workflows is NextFlow [58]. It supports deploying pipelines on a variety of execution platforms and provides support for managing workflow is dependencies through built-in support for Conda, Docker and Singularity.

IntelliOmics uses an alternative approach that relies on the queue message system (RabbitMQ [59]) and on the Celery [60] back-end to execute tasks in the background. Celery is a simple, flexible and reliable open-source distributed system with focus on real-time processing, while also supporting task scheduling. It can be easily extended with various plugins like Flower, which is a real-time monitor and web admin for Celery distributed task queue. We have integrated Celery within the pipeline module (Fig. 3)) to enable users to define and execute workflows, that are composed of new or predefined scripts and programs, that can run sequentially or in parallel. IntelliOmics is designed to work with virtually any analytical pipeline that can be easily incorporated into the system. User can choose between various tools and predefined guidelines and best practices [61], or create multiple alternative workflows e.g. for the commercial and research purposes, depending on licenses of the software. However,

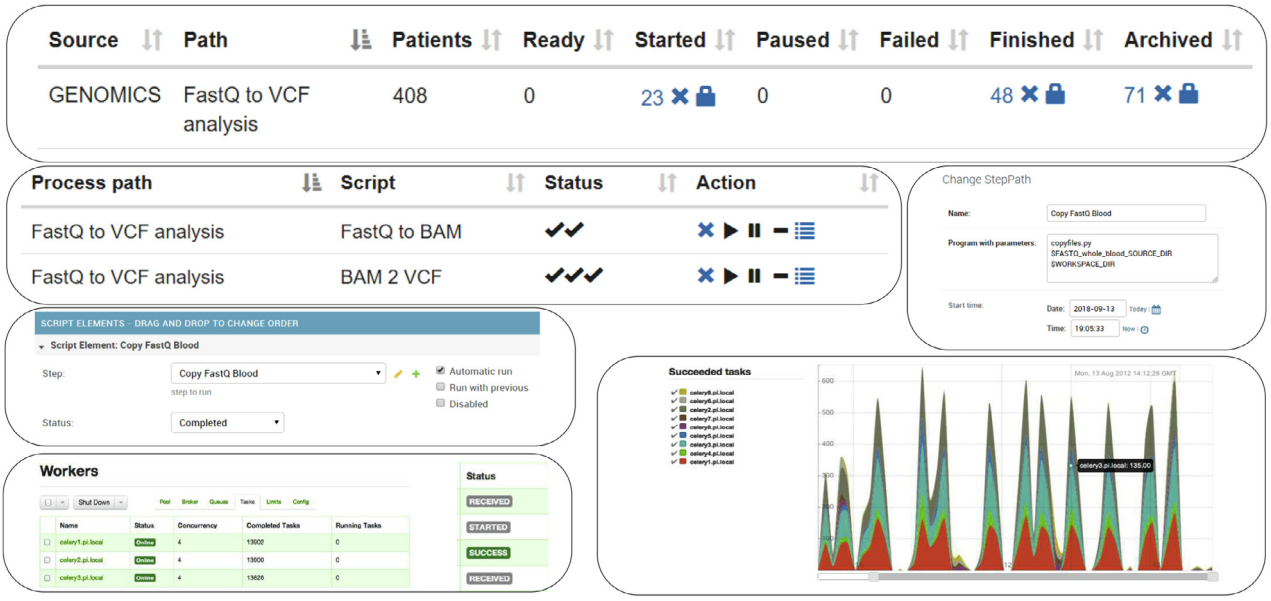

Fig. 3 – A bunch of example screenshots of the workflow tasks and scheduling.

user can easily use its own scripts, external algorithms or preferred external libraries to perform preprocessing and QC, adequate to specificity of the data and platform used for data acquisition. Next, he/she can monitor and manage those tasks which can be automatically distributed and queued on available resources.

Clinical data as well as a part of omic-data is stored in relational a SQL database whereas whole genome sequencing (WGS) data is stored in the Hive warehouse. The data layer is subject to version control. The system is capable of creating a snapshot of the currently stored data for a particular timestamp, so it is possible to reproduce the state of the database from a specific date or period. System administrators can manage and monitor Hadoop clusters with the free easy-to-use management web UI called Apache Ambari. Fig. 3 illustrates sample screenshots of IntelliOmics pipeline.

It is important to have a flexible workflow system for the data transformations and future analysis. Application of such pipeline can be especially useful for high-throughput genomics data where a transformation from raw files (e.g. FASTQ file format [62]) through Binary Alignment/MAP format (BAM) [63] to variant call format (VCF) [64] may be performed in various ways. In the literature, we may find some of the best practices including GATK workflows for whole genome sequencing, RNAseq analyses and other free and open source solutions like Gemini [65] for VCF annotation and exploration. A lot of effort should also be put into integrating clinical and multi-omics data [20]. Partial integration can be done automatically using workflows with external tools and scripts, however, a well designed system should allow users to create their own rules to select and integrate the data, as discussed in the following section.

2.3. *Data analysis*

Extracting useful information from the data and presenting them in a simple, accessible way is the most crucial part of any system for personalized medicine. Depending on the user type, system should provide interfaces to generate charts, statistics, reports as well as to apply data mining tools. It is also important, to allow collecting expert's knowledge (coming from e.g. physicians' practice, unpublished results and public databases) that can be recognized and used by the system. To improve the performance, computations could be separated into the client-side (charts, statistics) and the server-side (data mining, queries and reports). It can be realized with one of many available free frameworks based on the JavaScript language (IntelliOmics uses Angular [66] with JQuery).

The IntelliOmics data analysis part (see Fig. 1) is composed with a set of high-level interfaces to access the heterogeneous warehouse. The collected multi-omics data can be associated with expert's knowledge, e.g. clinical observations concerning a particular cancer diagnosis and treatment. It can be provided by the specialized visual editor carefully designed to work with multi-omics schema and clinical forms. All these elements can constitute so-called 'knowledge base' that allows the other groups of users (e.g. physicians) to explore and visualize collected data, without getting to know what is behind the user interface. In addition, embedded data-mining module can automatically extract hidden patterns within gathered data and generate a support for diagnostic and therapeutic purposes in personalized medicine.

2.3.1. *Knowledge base*

The idea of the knowledge base in the IntelliOmics is to allow the data and the expert's knowledge coming from the exter-

nal sources as well as from clinicians and other researchers to be integrated using visual editor. The knowledge is represented in a form of rules (called *observations*) evaluated (for a specific patient) to logical value: True or False and may concern clinical and multi-omics data. Rules can be use as filters to select and group patients with certain common characteristics, of which one part may be a hypothesis and the other - a diagnosis (both parts can be stored in the data warehouse). As soon as the hypothesis is positively verified (research stage), a filter consisting only of the first part can be used to find patients with a high probability of a given diagnosis (diagnostic stage). The system interface (see Fig. 4) allows using several types of observations:

- *simple* – built based on a single raw value/feature, coming directly from the warehouse, e.g.: *Old*: Age $>$ 70;
- *composite* – composed of one or more sub-observations, that performs certain operations on the data (logical, arithmetic etc.). This way the rules can be assembled from previously prepared, reusable units like: *OldMen* : *Old* AND *IsMen*;
- *function* – predefined (e.g. multiplication, square root) and grouping functions (average, sum, etc.) on selected one or more observations;
- *generic* – enable construction of a template observation, which parameter's values can be filled later from external source (e.g. csv file) or by the user. An example: *HighBloodPreasure*(#*X*): Blood preasure > #X, where #*X* can be given while using this generic observation (e.g. *HighBloodPreasure*(160)).

We enable hierarchical creation of the observations on the basis of previously defined components together with assigning them to multiple categories (and sub-categories). Just like building from blocks - users can select observations which differ in the level of detail and complexity. The idea is to let less technical users to just create and use simple and general high-level observations and 'hide' from them advanced and complex low-level definitions and parameters. It is up to the experts in particular field to define more advanced relations. In addition, there also exists information observations, which role is to control and limit the data returned by the filters.

Fig. 6 shows an example of a hierarchical observation with levels their view that are composed from the blocks of knowledge. Starting from the top, we see one of the examples of general observation related to the important mutations in the lung cancer. It is composed of several observations visible in the detailed view level. If the user does not want to check any particular conditions or does not know this particular field, he can simply use this observation and ignore the details. However, expert users can dig deeper and deeper (even to raw data), set desired conditions or even compose own observation from the chosen building blocks.

2.3.2. *Observation views*

Any integrated warehouse should allow a direct exploration by advanced users, e.g. bioinformaticians or computer scientists. Queries could be asked based on the clinical and multi-omics data using the SQL-like editor. Distributed SQL query engines (SQL-on-Hadoop) are designed to execute such queries on large datasets that are stored within big data warehouses. Notable SQL-on-Hadoop systems include Apache Impala [67], Apache Drill [68] an Facebook's Presto [69].

In IntelliOmics, we decided to incorporate Presto as a distributed SQL query engine because it was more stable and advanced when the system was designed. It supports interactive analytical queries against data sources of all sizes up to petabytes as well as previously defined set of queries created by knowledge base building blocks. IntelliOmics also offers more advanced observation views: one that is centered on a single patient, and one that is focused on a group of patients, whereas the user can define practically any group. In both interfaces a list of predefined observations can be selected and executed. The results can be viewed in form of tables and various charts, and the imagining data can be viewed and annotated using the DICOM module (Fig. 5). The web-based DICOM viewer performs an on-the-fly conversion of the raw imagining data. The module displays only the anonymized versions of the original images. Any unnecessary personal health information is therefore omitted. In addition, the system offers statistical analysis module that allows performing varied statistical tests and preparing definable reports which can be easily distributed or printed (e.g. as clinical reports in pdf).

2.3.3. *Exploration module*

The rapid development of large-scale technologies and an increased use of omics techniques in clinical practice is driv-

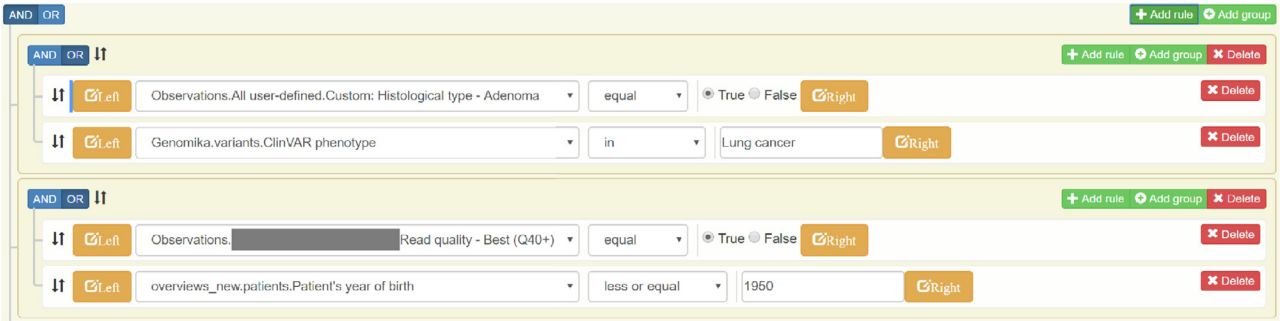

Fig. 4 – The interface to create a new block of knowledge.

Fig. 5 – Viewing the stored data using the SQL and observation views (top), and the results: tables, charts, images and reports (bottom).

ing researchers towards multi-omics [6,2]. This approach accelerates efficient prediction of the treatment response with substantially increased accuracy in personalized medicine. It also shows a strong correlation between molecular profiles and clinical outcomes in various types of cancers. A number of machine learning algorithms can be useful and reveal previously unrecognized molecular patterns associated with clinical phenotypes and can provide novel insights into the multi-omics data [20].

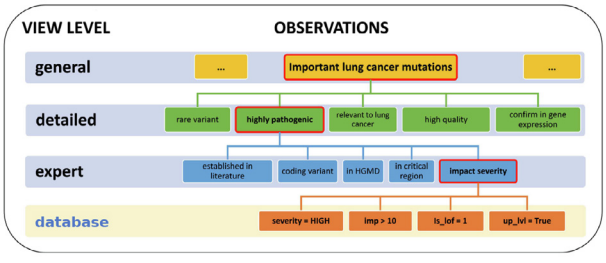

Fig. 6 – An example view of hierarchical observations.

Semi-automatic analyses of the gathered raw, transformed and filtered data can be performed with machine learning tools [4]. One of the most promising tools is the Spark's open-source distributed machine learning library called MLlib [70]. It provides efficient features for a wide range of learning settings, and includes several underlying statistical, optimization, and linear algebra primitives. Within the MLlib library, it is possible to perform fast and scalable large-scale data analysis including classification, regression, collaborative filtering, clustering, and dimensionality reduction. Additionally, Spark MLlib also provides easy-to-use APIs that enable deep learning, which is currently attracting interest in health informatics [5]. However, Spark's MLib is not really user friendly, therefore, additional high-level interfaces should be build upon it to make it available for the less technical users.

3. Results

The IntelliOmics is created within the framework of a medium-scale medical project [43] to support molecular diagnosis and therapy individualization in oncology. The project is called MOBIT and is focused on the non-small cell lung cancer (NSCLC) which is the most common type of lung cancer. It accounts for 85% of all lung cancer cases and include subtypes like: squamous-cell carcinoma, adenocarcinoma, and large cell carcinoma. Usually symptoms of NSCLC do not appear until the disease is already at an advanced stage. However, clinical studies have shown that individualized molecular targeted therapies increase survival time and improve patients' quality of life [48]. The result of the development phase of the project is a design of unique software for the omic- and clinical data acquisition, integration and analysis for use in the implementation of individualized therapy. The IntelliOmics system has been developed and implemented in close collaboration and supervision with various research centres and university clinical hospitals.

3.1. Hardware architecture

The current architecture for IntelliOmics is illustrated in the Fig. 7. It uses 22 Dell PowerEdge R440 servers with 8-core Intel Xeon Silver 4410: 2 managing servers - AccessNode (64 GB RAM, 2x480 GB SSD) and NameNode (64 GB RAM, 2x480 GB SSD, 2x2TB HDD); and 20 working servers DataNode (32 GB RAM, 2x2TB HDD). There is 60 TB available for data in a distributed HDFS file system. That relatively inexpensive computer cluster costed only 50 000$ but its performance is enough for efficient data analysis. For the most computing-demanding tasks involving whole genome sequencing data - a search for a single mutation for a particular patient is less than a quarter of second. To filter and scan for all important mutations that activate EGFR gene (over 40 mutations) it takes 7 s for a single patient, and less than 2 min for all the patients. Proposed system infrastructure is cost-effective and can be easily scaled to analyze larger volumes of data.

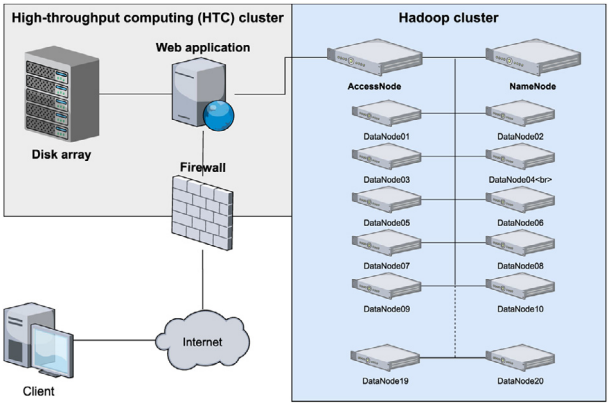

Fig. 7 – Hardware architecture of the IntelliOmics system.

3.2. Collected data and performed analysis

The highest standards of the gathered data and their processing are the result of: reliable selection of patients; collection of a high-quality biological materials; the most advanced research equipment for high-throughput studies; experience in analysis of large-scale data; access to the know-how in the field of biobanking and the hybrid PET/MRI system. In the Mobit project we integrated and explored collected data (see Section 2.1) from over a hundred of patients suffering non-small cell lung cancer and control group. The data includes personal and clinical information, imaging data and the results from biological materials.

The data collections for a typical patient is as follows:

- patient or nurse fills up the survey forms with various types of data describing the patient, disease and medical examination. Additional results from biological materials, DXA (Dual Energy X-ray Absorptiometry), inBody, retinal scan, exercise tests (e.g. treadmill) or even pathologist's report are imported automatically or manually to the system to previously defined forms.
- the genomics and transcriptiomics data is automatically transferred to the storage matrix server. As each file has unique, specified name composed of patientID, project type, data source, data type and timestamp, the system recognize and semi-automatically imports the data.
- the whole genome sequencing (WGS) is performed with Illumina HiSeq4000 instrument on a patient cancer tissue and blood. We focused on the cancer genomic alternations, including point mutations, small insertions or deletions, copy number alterations, somatic and structural variations. The use of high-coverage transcriptomics, allows to not only quantify gene expression profiles, but also detect alternative splicing, editing and fusion of transcript.
- metabolomics analyses have been performed by use LCQTOF-MS and peptide samples were analyzed on a the Autoflex-ToF/ToF mass spectrometer. Serum and tumor tissue samples of patients with NSCLC and controls were fingerprinted. Leukocyte and urine fingerprinting methodology have been applied as well. The exported results of these analysis in a csv format are next imported into the system.
- the PET/MRI and CT images from the patients with NSCLC before tumor resection, and after surgery are also imported to the system in a DICOM format. Currently, we do not examine the raw images but only their description in a form of a defined survey form.

Next step is the data processing and integration. Designed and implemented analytical pipelines for the next-generation sequencing data analyses include the following technologies, their intersections and downstream analyses: RNAseq, smallRNAseq, methylated DNA and genome sequencing. Obtained data has been used to detect sequence and structural polymorphisms that was further correlated with the phenotypical data (drug resistance, prognosis, etc.) as well

as with results from other technologies (e.g. gene expression levels and DNA methylation). As a result, we were able to detect genes and isoforms that were up and down regulated depending on tumor stage and phenotype. IntelliOmics is designed to work with any workflow, scripts or external tools as it monitors and controls when, where and what is running. During the project we often changed, tested and adapted different tools due to their very dynamic development. The same applies to databases e.g. for annotation and even reference genomes as depending on user requirements two were actively used: GRCh38 and GRCh37. Therefore, it is important to let bioinformaticians choose the best according to their tools, transformations and databases.

In our experiments we follow the GATK protocols and workflows from NextFlow [58]. The most time and resource consuming task was the WGS analysis which takes a minimum of one-two days to complete per sample. Hopefully, we could analyze a good number of samples in parallel. As for the other omics the calculations took no more than few hours to finish.

3.3. *Knowledge base*

The core of the MOBIT team consortium consists of people from both research (bio-medicine, bioinformatics and computer scientists) and hospital (physicians and clinicians). Thanks to the team experience and scientific background we have managed to extend the knowledge base with several hundreds of observations specific to the non-small lung cancer (including lists of mutations, genes of special interest and potential drivers); several dozen of study scenarios that check various patient attributes stored in the warehouse; thousands of filters and multi-level condition blocks.

Below three examples of knowledge extraction within IntelliOmics for the non-small lung cancer are reported. Please consider this as a brief example of the system's capabilities rather than an actual medical analysis.

3.3.1. *Potential therapy with Tagrisso*

The domain knowledge in a form of recommendation for application of Tagrisso (trade name of osimertinib) was defined as a tree of observations (Fig. 8). It contains several preconditions from multiple data sources:

- patient's sex and clinical history of his treatment (e.g. taken medicament to avoid potential interactions) – from questionnaires;
- kidney, liver and blood system parameters – from the blood morphology examination;
- presence of specific mutations – from genomics data;
- cancer type – from a pathologist's report.

The final tree was constructed from previously defined observations joined by AND and OR operators. Such a modular character of construction allows easy reusing of the observations as building blocks. The tree can be interactively validated – after expanding any of its branches each observation on every level can be evaluated independently. The anal-

Fig. 8 – Interactive observation tree for Tagrisso recommendation allowing examination of the results for a patient group.

ysis can be run and its results may be presented for a single patient (positive or negative) or any group of patients. The group is specified via mechanism of a filtering observation.

Some nodes (i.e. mutation and drug interaction ones) were constructed in a generic way (*generic observations*), where actual parameter values were replaced with symbolic parameters, which were set further within specialized binder wrappers. Here is an example of EGFR gene mutation's parameters:

```
{#var_type = 'snp',          #gene = 'EGFR',
#vcf_start = 55249070,       #vcf_end = 55249071,
#ref = 'C',                  #alt = 'T',
#var_sub_type = 'ts',        #aa_change = 'T790M',
#chrom = 'chr7'}
```

Depending on the mutation type, the parameters involve the variant position, change and additional annotations [65] gathered by the tools like Annovar, SnpEff, VEP etc.

3.3.2. *Novel mutations connected with non-small lung cancer*

Within IntelliOmics, we have created various multi-omics knowledge observations using the interface presented in Fig. 6. For an over 3 million gene variants (per patient) we filtered around 200 000 high quality variants (based on sequencing quality and quantitative metrics) with over 90 000 alterations in the gene coding regions. Next, we filtered highly pathogenic variants, excluding existing ones in popular gene datasets (COSMIC, ClinVar, HGMD, OMIM), and focused on changes that appeared in genes related to non-small lung cancer. Several thousands of potential mutations were then confronted with gene fusions and microRNA control site and the main attention has been focused on somatic mutations. With such an approach, we managed to find no more than dozen potentially highly pathogenic novel mutations for both group of patients: adenoma and squamous cell carcinoma. Currently, our multidisciplinary research team is investigating our findings to validate their biological meaning. The next possible step is to scan for pathways altered by the gene mutations (starting from SNP mutations, through genes expressions, protein coding till metabolomic data).

3.3.3. *Data exploration with MLlib*

Sample analysis was performed with MLlib in order to recognize the type of non-small lung cancer: adenoma (AD) and squamous cell carcinoma (SCC). Among the available algorithms the decision trees (DTs) [71] were selected. DTs have proven to be successful in biomedical analysis [71] and their prediction models are easy to interpret. We have performed the experiments with 10-fold stratified cross-validation using four types of data: (a) 4 basic clinical information: sex, age, pack-year for smokers and the stage of disease; (b) over 500 metabolites from plasma with positive polarization and reverse-phase (RP) chromatographic separations in metabolomic studies; (c) 28 selected microRNA variables and (d) 16 selected single nucleotide polymorhisms (SNPs) related to the non-small lung cancer. Since not all the patients had a complete set of tests (see Fig. 9), we have performed five variants of experiments, depending of the available data: (a)+(b), (a)+(c), (a)+(b)+(c), (a)+(c)+(d), (a)+(b)+(c)+(d).

Results enclosed in Fig. 10 show two examples of decision trees and their classification accuracies. In both cases the generated model considers omic data more often to make the predictions and skipped clinical and/or microRNA information (see Fig. 10). The possible reason is the high ratio of metabolomic features to relatively low number of patients. This problem often refers as the curse of dimensionality as reclassification quality results is much higher than the classification quality on an independent set. However, lower prediction quality for the variants without metabolomics data (Fig. 10-right) suggests that found metabolites may, in fact, have important discriminative potential. Both trees are very simple, provide logical reasoning and may actually help in understanding and identifying relationships between specific features and improve biomarker discovery.

This way, IntelliOmics is able to generate many predictive models based on different combinations of multi-omics sets. In a sense, our system could act as artificial panel/medical review board, where each expert (induced decision tree) presents conclusions based on different types of data. One could look at the patient from different perspectives and the decision support could be prepared semi-automatically.

Our system provides the possibility of using different models, like Random Forests (RF) [72] which average multiple decision trees and are popular in medical data mining. Performed experiments show that RF managed to achieve high reclassification accuracy (97.8% Fig. 10-left and 95.5% Fig. 10-right) but relatively much lower classification accuracy (56.3% and 64.3% respectively). However, when modeling is aimed at understanding basic environmental processes, such methods are not so useful because they generate more complex and less understandable models.

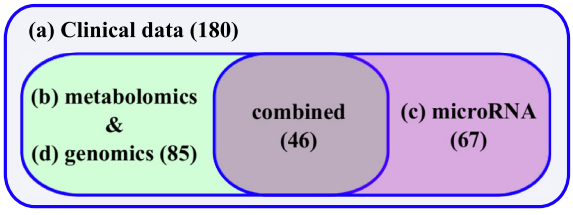

Fig. 9 – The number of patients who have (a) clinical, (b) metabolic, (c) microRNA and (d) genomic tests performed.

However, to interpret information contained within the various attribute types and get some biological understanding, a much larger group of patients is needed. Additional validation by an independent cohort is also required to draw any conclusion or claim biological or clinical evidence. Finally, a care should be taken to ensure that the histological assignment of the patients is as accurate as possible (e.g., evaluation by multiple independent pathologists).

4. Discussion

The closest thing to our system is the Databricks solution [41], which was the direct inspiration for our Inteliomics system. This one of the largest frameworks for massive genomics analysis is based on the best open-source solutions intelligently wrapped into a unified data web-based platform. The main difference is that the service is available on a cloud hosting platform and cannot be run in private data-centers or on-premise clusters. This raises questions about costs, which can be very high, especially for genomics data that are not only large in volume but also require a lot of preprocessing and transformations. The biggest positive surprise with databricks has been the speed of processing genomic data. Running standard NGS pipelines involing alignment and variant calling were about 20–50 times faster than the calculations performed on our IntelliOmics software servers. As an excuse, we must add that our computing servers were not among the fastest. The main drawback of Databricks is that it loses some of the flexibility available in open-source components that come with it. We were unable to run custom scripts or workflows that use alternative algorithms. Finally, the Databricks is targeted at computational biologists and bioinformaticians, and is limited to genomic data only, so clinical or other omics cannot be used in the analysis.

Another powerful commercial product is Ingenuity Variant and Pathway Analysis (IPA) [47] from QIAGEN Bioinformatics [46]. It is the leading pathway analysis application among the life science research community for analysis, integration and interpretation of data from –omics experiments. Like Databricks it is a web-based application, but unlike Databricks it allows multi-omics analysis. Similar calculations takes much more time on IPA, but can be done not only in cloud but also on the CLC genomic server or even on a personal computer. The system is a closed box and does not allow you to run any other algorithms or analysis other than those implemented in the software. However, the number of possible automatic analysis and solutions along with the extensive built-in knowledge base is impressive, but this comes with a high license cost. Finally, as in Databricks, importing more complex clinical data or additional user knowledge is limited.

Currently, the most popular free solution for computational research is a software framework called Galaxy [45]. It is an open, web-based platform for accessible, reproducible and transparent biomedical analysis, available on public servers, in the cloud or locally. Galaxy allows bioinformaticians

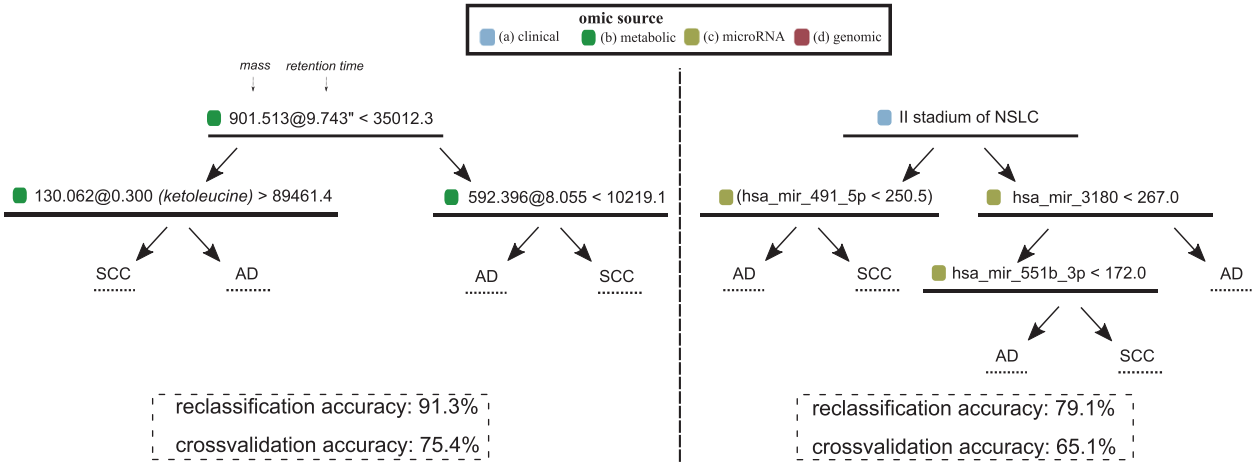

Fig. 10 – Examples of decision trees generated for the integrated (a) clinical, (b) metabolic , (c) microRNA and (d) genomic data. Left tree was induced with (a)+(b), (a)+(b)+(c) and (a)+(b)+(c)+(d) whereas right tree with (a)+(c) and (a)+(c)+(d) data.

and researchers to define and run workflows that include everything from data loading, transformation, processing to displaying analysis results. The system is largely defined by tools and extensions proposed by the community. Although originally developed for genomic research, it is largely domain agnostic and is now used as a general bioinformatics workflow management system. Recently, a number of specialized plugins or add-ons have emerged to perform complete and integrated analysis for multi-omics extensions [50] and visualizations [49]. The Galaxy framework is great as long as it comes down to running specific analyses by scientific communities. However, it may not work well as a complete solution for running a medical project that will benefit not only scientists, but also nurses or clinicians.

In Tables 1 and 2 we have shown reviews that address a much larger number of omics-related software, tools, and other approaches, which already number in the hundreds. New algorithms, updates, pipelines, and systems appear every day, but most of these new developments are of interest primarily to bioinformaticians and researchers. Broader user groups associated with the medical centers or local research projects seek for more comprehensive user-friendly services that help to storage, manage and analysis various types of omics, imaging and clinical data. Unfortunately, it is impossible to create a single system that will meet all user needs. However, we have shown that with some effort it is possible to create end-to-end software tailored to the needs of a wide audience (see Table 3) using free, well-designed and maintained components with an open license for both personal and commercial use.

There are several issues and concerns to consider before creating your own system. First of all, unless you want to write your own solutions from scratch, you are limited to existing solutions that are open source or with easily accessible API (application programming interface). In this work we proposed those that we believe are the best suited to for medium or even large medical projects. Some of the proposed open source systems are already successfully used in commercial products such as Databricks. While we wanted to show an essential list of tools, there are many other specialized approaches that may be applicable, such as NLP (Natural Language Processing) for sentiment analysis based on physician descriptions in clinical data or much more advanced data mining libraries besides MLLib. When it comes to raw data processing, transformation and partial integration, we found the approach used in Galaxy software to be the most suitable. In this field, the dynamics of the emergence of new tools, libraries, databases (e.g. for annotation) is so high that this part should be as flexible as possible. Therefore, bioinformaticians can simply run their scripts or pipelines under the supervision of a system that monitors and allocates resources. Finally, in terms of presentation of results and overall user-friendliness, we found IPA from QIAGEN to be the most attractive. However, it all comes down to the needs of the specific project and the people that will use that software. As a proof of concept we developed IntelliOmics for amid-scale medical project [43] focused on precision medicine for lung cancer patients which was described in detail in previous sections.

The proposed system can also be the basis for practical clinical applications. Once we have populated the system's database with patient information, it is necessary to analyze the collected data and look for patterns or relationships. The next natural step is to apply derived and validated knowledge on new patients or groups of patients from the hospital e.g. in clinical trials. In the case of the MOBIT project [43], we worked on a lung cancer management tool incorporating PET/MRI, genomics and biomarkers as a decision-making tool for treatment selection. This software is designed based on data from a research model (patients diagnosed with lung cancer undergoing surgical treatment) and data from a designated patient group, i.e., patients with suspected NSCLC or at high risk. Newly obtained biomarkers can help clinicians to diagnose cancer in high-risk patients before the appearance of visible lung lesions. Another practical clinical application of the collected information is the individualization of treatment selection of cancer patients (chemotherapy). By cyclically collecting information about the drugs administered to

patients (type, dose, time of administration) and their health status, it is possible to build a model suggesting the best potential treatment for future patients. Thus, having an integrated knowledge base with tools for their analysis that also include elements of artificial intelligence can make a huge contribution to personalized medicine.

5. Conclusions

Building a private, standalone system for multi-omics data analysis towards personalized medicine may look as a very difficult task. However, this process can be greatly simplified and accelerated by using well-designed and maintained components with open source license for both personal and commercial use. Our proposition aims to serve some guidelines how to create your own local end-to-end service for medium size or even a large-scale projects. We have picked what we believe to be the best features from existing systems both free and commercial to integrate them into a single system. The overall framework is modeled after commercial product Databricks which was constituted on the Hadoop ecosystem which is efficient for the big data analysis. In terms of flexibility of data transformation, processing and integration we were inspired by the Galaxy system in which the user can simply run its own scripts or tools. As a result, we present advantages and disadvantages of dozen tools along with the recipes how to include them in building your own system.

To validate this guidelines, we have developed the IntelliOmics. It is free, personalized and flexible solution for multi-omics data analysis is capable of integrating quantitative omics data and builds prediction models for cancer phenotypes, making the state-of-the-art machine-learning methods accessible to researchers of all backgrounds. The results are very promising and there are already a number of projects related to other cancers as well as e.g. diabetes for which such a designed system is applicable. Currently, IntelliOmics is at the last stage of the clinical interfaces completion, so the system can display aggregated knowledge in user-friendly formats. As a result, physicians, non-technical laboratory personnel, medical assistants, data scientists and research coordinators, and other end-users can enter data, access information, mine the datasets and understand the output. We believe that its total cost (hardware + software development) is relatively small compared to the benefits and possibilities it offers.

The roadblocks and future works include ethical, legal, and logistical concerns. In addition, ensuring data security and protection of patient rights while simultaneously facilitating standardization is a principal rule in the public support maintenance. Finally, we would like to extend our pipelines with the radiomics-based analysis of the medical imaging [73]. This approach is becoming increasingly important in cancer studies [74] and personalized medicine [75].

CRediT authorship contribution statement

Daniel Reska: Software, Conceptualization, Writing - original draft, Visualization, Formal analysis, Validation, Investigation. **Marcin Czajkowski:** Conceptualization, Writing - original draft, Methodology, Formal analysis, Data curation. **Krzysztof Jurczuk:** Conceptualization, Writing - original draft, Visualization, Methodology. **Cezary Boldak:** Software, Writing - original draft, Visualization, Formal analysis. **Wojciech Kwedlo:** Software, Data curation, Writing - original draft, Visualization, Formal analysis, Validation. **Witold Bauer:** Data curation, Writing - review & editing, Methodology, Investigation. **Jolanta Koszelew:** Project administration, Resources, Investigation, Validation. **Marek Kretowski:** Funding acquisition, Supervision, Conceptualization, Writing - review & editing, Resources, Investigation, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The project funded by the National Centre of Research and Development in the framework Programme “Prevention practices and treatment of civilization diseases” - STRATEGMED (STRATEGMED2/266484/2/NCBR/2015).

Acknowledgments

The authors would like to thank for collaboration all MOBIT team researches and especially to: Jacek Niklinski, Miroslaw Kwasniewski and Michal Ciborowski from Medical University of Bialystok and Konrad Kozlowski from Bialystok University of Technology.

REFERENCES

[1] Pashazadeh A, Navimipour NJ. Big data handling mechanisms in the healthcare applications: a comprehensive and systematic literature review. J Biomed Inform 2018;82:47–62.
[2] Lin MC, Iqbal U, Li YC. AI in medicine: big data remains a challenge. Comput Methods Programs Biomed 2018;164.
[3] Tang KJW et al. Artificial intelligence and machine learning in emergency medicine. Biocybern Biomed Eng 2021;41 (1):156–72.
[4] Mirza B, Wang W, et al. Machine learning and integrative analysis of biomedical big data. Genes (Basel) 2019;10(2):87.
[5] Wu PY et al. Omic and electronic health record big data analytics for precision medicine. IEEE Trans Bio-medical Eng 2017;64(2):263–73.
[6] Ashley EA. Towards precision medicine. Nat Rev Genet 2016;17(9):507–22.
[7] Tran B et al. Cancer genomics: technology, discovery, and translation. J Clin Oncol 2012;30(10):647–60.
[8] Thapa C, Camtepe S. Precision health data: requirements, challenges and existing techniques for data security and privacy. Comput Biol Med 2021;129 104130.
[9] Kalina J, Matonoha C. A sparse pair-preserving centroid-based supervised learning method for high-dimensional

biomedical data or images. Biocybern Biomed Eng 2020;40 (2):774–86.
[10] Viceconi M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. IEEE J Biomed Health Inform 2015;19(4):1209–15.
[11] Momeni Z et al. A survey on single and multi omics data mining methods in cancer data classification. J Biomed Inform 2020;107 103466.
[12] Shahid AH, Singh MP. Computational intelligence techniques for medical diagnosis and prognosis: Problems and current developments. Biocybern Biomed Eng 2019;39(3):638–72.
[13] Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, Wishart D. Systems biology and multi-omics integration: viewpoints from the metabolomics research community. Metabolites 2019;9(4):76.
[14] de Anda-Jáuregui G, Hernández-Lemus E. Computational oncology in the multi-omics era: state of the art. Front Oncol 2020;10(423).
[15] Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet 2017;8(84).
[16] Computational Data Analysis Workflow Systems. https://s.apache.org/existing-workflow-systems.
[17] Chung R, Kang C. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. GigaScience 2019;8:5.
[18] Zanfardino M et al. MuSA: a graphical user interface for multi-OMICs data integration in radiogenomic studies. Sci Rep 2021;11:1550.
[19] Misra BB, Langefeld C, Olivier M, Cox LA. Integrated omics: tools, advances and future approaches. J Mol Endocrinol 2019;62(1).
[20] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights 2020;14. 1177932219899051.
[21] Labory J et al. Multi-omics approaches to improve mitochondrial disease diagnosis: challenges, advances, and perspectives. Front Mol Biosci 2020;7 590842.
[22] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res 2018;46(20):10546–62.
[23] Lee B et al. Heterogeneous multi-layered network model for omics data integration and analysis. Front Genet 2020;10:1381.
[24] Zeeshan A, Khalid M, Saman Z, XinQi D. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database 2020.
[25] Cirillo D, Valencia A. Big data analytics for personalized medicine. Curr Opin Biotechnol 2019;58:161–7.
[26] Tong L et al. Integrating multi-omics data by learning modality invariant representations for improved prediction of overall survival of cancer. Methods 2021;189:74–85.
[27] Gambardella V et al. Personalized Medicine: Recent Progress in Cancer Therapy. Cancers (Basel) 2020;12(4).
[28] Su M et al. Proteomics, Personalized Medicine and Cancer. Cancer 2021;13:2512.
[29] Morello G et al. From multi-omics approaches to precision medicine in amyotrophic lateral sclerosis. Front Neurosci 2020;14 577755.
[30] Hou X et al. The multi-omics architecture of juvenile idiopathic arthritis. Cells 2020;10(10). 2301.
[31] Song JW et al. Omics-driven systems interrogation of metabolic dysregulation in COVID-19 pathogenesis. Cell MeTable 2020;32(2):188–202.
[32] Miki D et al. Hepatocellular carcinoma: towards personalized medicine. Cancer Sci. 2012;103(5):846–50.
[33] Rivenbark AG et al. Molecular and cellular heterogeneity in breast cancer: challenges for personalized medicine. Am J Pathol 2013;183(4):1113–24.
[34] Krzyszczyk P et al. The growing role of precision and personalized medicine for cancer treatment. Technology (Singap World Sci) 2018;6(3–4):79–100.
[35] Frohlich H et al. From hype to reality: data science enabling personalized medicine. BMC Med 2018;16:150.
[36] Couri T et al. Goals and targets for personalized therapy for HCC. Hepatol Int 2019;13(2):125–37.
[37] Zanfardino M et al. Bringing radiomics into a multi-omics framework for a comprehensive genotype–phenotype characterization of oncological diseases. J Transl Med 2019;17:337.
[38] Zeng Y et al. Bioinformatics analysis of multi-omics data identifying molecular biomarker candidates and epigenetically regulatory targets associated with retinoblastoma. Medicine (Baltimore) 2020;99(47):e23314.
[39] Lawal B et al. Multi-omics data analysis of gene expressions and alterations, cancer-associated fibroblast and immune infiltrations, reveals the onco-immune prognostic relevance of STAT3/CDK2/4/6 in human malignancies. Cancers (basel) 2021;13(5):954.
[40] Xie F et al. Three-dimensional bio-printing of primary human hepatocellular carcinoma for personalized medicine. Biomaterials 2021;265 120416.
[41] Etaati L. Azure databricks. Mach Learn Microsoft Technol 2019;159–171.
[42] Zaharia M et al. Apache Spark: a unified engine for big data processing. Commun ACM 2016;59(11):56–65.
[43] Niklinski J et al. Systematic biobanking, novel imaging techniques, and advanced molecular analysis for precise tumor diagnosis and therapy: the Polish MOBIT project. Adv Med Sci 2017;62(2):405–13.
[44] Silva BN, Khan M, Han K. Internet of things: a comprehensive review of enabling technologies, architecture, and challenges. IEEE Tech Rev 2018;35(2):205–20.
[45] Afgan E et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 2018;46(1):537–44.
[46] QIAGEN Inc.,http://qiagenbioinformatics.com .
[47] Yu J, Gu X, Yi S. Ingenuity pathway analysis of gene expression profiles in distal nerve stump following nerve injury: insights into wallerian degeneration. Front Cell Neurosci 2016;10:274.
[48] Nema R, Shrivastava A, Kumar A. Prognostic role of lipid phosphate phosphatases in non-smoker, lung adenocarcinoma patients. Comput Biol Med 2021;129 104141.
[49] McGowan T et al. An extensible Galaxy plug-in for multi-omics data visualization and exploration. GigaScience 2020;9:4.
[50] Mehta S et al. Precursor intensity-based label-free quantification software tools for proteomic and multi-omic analysis within the galaxy platform. Proteomes 2020;8(3):15.
[51] Patil S, Majumdar B, Awan KH, Sarode GS, Sarode SC, Gadbail AR, Gondivkar S. Cancer oriented biobanks: a comprehensive review. Oncol Rev 2018;12(1):357.
[52] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18:83.
[53] Paglialonga A, Lugo A, Santoro E. An overview on the emerging area of identification, characterization, and assessment of health apps. J Biomed Inform 2018;83:97–102.
[54] Ahmed AJ. Django Project Blueprints. UK: Packt Publishing; 2016.
[55] Leipzig J. A review of bioinformatic pipeline frameworks. Briefings Bioinform 2017;18(3):530–6.

[56] Erraissi A, Belangour A, Tragha A. Digging into hadoop-based big data architectures. Int J Comput Sci 2017;14(6):52–9.
[57] Camacho-Rodríguez J et al. Apache Hive: From MapReduce to enterprise-grade big data warehousing. ACM SIGMOD 2019;1773–1786.
[58] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017;35(4):316–9.
[59] RabbitMQ URL:http://www.rabbitmq.com/ .
[60] Celery URL:http://www.celeryproject.org/ .
[61] Castel SE et al. Tools and best practices for data processing in allelic expression analysis. Genome Biol 2015;16:195.
[62] Cock PJA et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 2010;38(6):1767–71.
[63] Li H et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078–9.
[64] Danecek P et al. The variant call format and VCFtools. Bioinformatics 2011;27(15):2156–8.
[65] Akgun M, Demirci H. VCF-Explorer: filtering and analysing whole genome VCF files. Bioinformatics 2017;33(21):3468–70.
[66] Freeman A. Putting Angular in Context. Pro Angular. Apress, Berkeley, CA; 2017.
[67] Bittorf M et al. Impala: a modern, open-source SQL engine for Hadoop. In: Proceedings of the 7th Biennial Conference on Innovative Data Systems Research, CIDR '15.
[68] Hausenblas M, Nadeau J. Apache Drill: interactive ad-hoc analysis at scale. Big Data 2013;1(2):100–4.
[69] Sethi, R. et al.: Presto: SQL on Everything. ICDE'35 1802-1813 (2019).
[70] Meng X, Bradley J, Yavuz B, Sparks E, et al. MLlib: machine learning in Apache Spark. J Mach Learn Res 2016;17 (1):1235–41.
[71] do Nascimento PM, Medeiros IG, Falcão RM, et al. A decision tree to improve identification of pathogenic mutations in clinical practice. BMC Med Inform Decis Mak 2020;20(52).
[72] Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics 2012;99(6):323–9.
[73] Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016;278(2):563–77.
[74] Thawani R et al. Radiomics and radiogenomics in lung cancer: a review for the clinician. Lung Cancer 2018;115:34–41.
[75] Lambin P et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14(12):749–62.