A Service for Flexible Management and Analysis of Heterogeneous Clinical Data

Sandro Hurtado[1,2,3][0000−0003−0990−480X], José García-Nieto[1,2,3][0000−0003−2985−3480], and Ismael Navas-Delgado[1,2,3][0000−0001−7819−5416]

[1] Dept. de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Spain
[2] ITIS Software, Arquitecto Francisco Peñalosa 18, 29071, Málaga, Spain
[3] Biomedical Research Institute of Málaga (IBIMA), Spain sandrohr@uma.es

Abstract. This paper describes FIMED 2.0, a Service for Flexible Management and Analysis of Heterogeneous Clinical Data. This software tool allows flexible clinical data management from multiple trials, which can help to improve the quality of clinical data and ease in clinical trials. The proposed service has been developed on top of a NoSQL Database (MongoDB), which allows for collecting and integrating clinical data in dynamic and incremental schemes based on their needs and clinical research requirements. Building upon our experiences with Flexible Management of Biomedical Data (FIMED), we have developed this new version of the tool aiming not only at replicating the former one but also including further gene regulatory network analysis and data visualization oriented to annotate gene functionality and identify hub genes. This version allows the practitioner to use four different network construction methods such as data assimilation, linear interpolation, tree-based ensemble or Gradient Boosting Machine regression. You may find a free version of this tool on the web at https://khaos.uma.es/fimedV2. A demo user account has been created to provide user demonstration, *"iwbbio"*, using the password *"demo"*. A real-world use case for a clinical assay in Melanoma disease is also included in this demo, which has been indeed anonymized.

Keywords: Clinical Research · Clinical Trial Management Systems · NoSQL Database · Gene Expression Data Analysis · Gene Regulatory Network Inference.

1 Introduction

Next-generation sequencing (NGS) has improved clinical genetics by allowing researchers to investigate hundreds of genes at once compared to conventional Sanger sequencing [17]. To conduct more in-depth analysis, clinical researchers integrate these data with other patients' clinical and personal information, such as electronic health records, habits, inheritance, and environmental factors [16]. In this regard, advanced data management and analysis systems in clinical study and personalised medicine have been developed over the last few years. Even with the tremendous advancement in NGS technology and bioinformatics software

tools, more improvements are required to deal with complicated and genetically heterogeneous diseases.

There are many software packages already available for clinical data management software and some of them are freely available to clinicians as found in [2, 6, 1, 9, 15, 3, 9, 15, 3]. The majority of these clinical data management systems are web-based platforms adapted to the requirements of a particular clinical trial, where clinicians have to design electronic Case Report Forms (eCRF) in any spreadsheet software and upload them through the user interface. Due to the heterogeneity of clinical data and the possibility for changes in different studies, clinicians must continually load forms into the system. In this sense, these systems lack flexibility and extensibility due to the database limitations in terms of development.

In addition, almost none of these systems allow the analysis of patients' clinical data to diagnose clinical diseases, so they are limited in filling the gap among bioinformatics, molecular geneticists and clinicians.

In this paper, we present FIMED 2.0, a software tool for the flexible management and analysis of clinical data. Our motivation for developing FIMED 2.0 stems from our experience with FIMED [7]. Our goal is provide users with new functionalities in order to preform new and more accurate analysis. In this new version, we place our interest in the analysis of Gene regulatory networks (GRNs) inference incorporating new GRNs algorithms for the sake of a principled comparison among GRNs gene network reconstructions. Also, an ensemble of GRNs has been proposed based on a voting system to allow users rank the most important gene interactions (top-k genes/edges) between the similar outputs of a set of GRNs. So, this can indicate the gene pairs that are most important in the regulatory process. Moreover, visualization tools have been added to the tool to provide users with a deep insight into the networks through a better graphic plotting. As a result, the primary goal would be to establish links between genes that are expressed similarly, which would lead to the discovery of novel therapeutic targets or biomarkers for the patient's expected treatment progression.

We demonstrated the advantages of the new functionalities of FIMED 2.0 in a practical use case using real expression data from metastatic Melanoma patients used in previous works [14, 7].

The remaining of this article is organized as follows. Section 2 describes the system architecture, detailing the gene expression pre-processing and gene regulatory inference methodology. In Section 3, a use case is reported, to illustrate the usability of the proposed service, and finally, in Section 4, the key conclusions and future directions of effort.

2 System Architecture

FIMED 2.0 is an extension of FIMED that internally implements a workflow as depicted in Figure 1, which consists of several phases: data collection, integration, clinical data analysis and data visualization. Thanks to the web interface, the user is guided through this workflow, so internal data mappings and adaptations

are automatically conducted. This section summarises the architecture of this tool, making special emphasis on the new elements included in this extension.

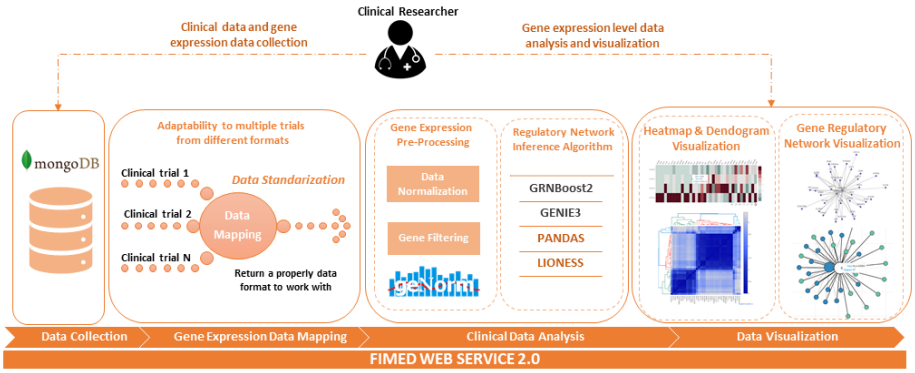

Fig. 1: FIMED 2.0 Workflow is made up of several phases: (I) Data collection, (II) Mapping data which guarantee the adaptability to multiple trials, (III) Clinical Data analysis through several algorithms and (IV) Data Visualization

2.1 Data Collection

The core of the data integration and management is a MongoDB database that provides a flexible way of dealing with clinical information such as gene expression data. The database schema is shown in the JSON Code Snipped 1.1. The database is organised into one collection of users. So, each user is represented as a BSON document. In this schema, the users are the clinicians, and so, each data entry includes the list of patients related to them. This means usually that for each clinician we have the patients participating in a clinical trial. The information associated with each patient is stored together, including general information for statistical studies, as well as any additional files related to the clinical trial (gene expression data, informed consent, blood sample reports).

However, end-users are not expected to directly interact with MongoDB database. Thus, FIMED provided a Web Graphical User Interface for easing the management of the clinical trial information. FIMED 2.0 has extended not only the list of available algorithms but also the visual tools provided to the end-users.

The Web GUI provides a dynamic visualization of the information that can modify the internal database schema depending on the clinical trial needs. Thus, if a user needs to include a new data field in the clinical trial, they only have to add it in the user interface and this will automatically update the implicit schema of the MongoDB database.

This tool also takes into account the previous experience of clinicians, so the data fields of previous patients is obtained and provided at the time of

adding new patients. This automatic process accelerates the data collection process reducing the database maintenance procedures done by technicians to the minimum.

Code Snippet 1.1: Core JSON Schema. It is the initial document structure from which the database adds new items and updates existing ones progressively.

```
{   "_id": <ObjectId()>,
    "Name": <String>,
    "Surname": <String>,
    "Password": <String>,
    "Patients": [{
            "_id": <ObjectId()>,
            "_patientInformation": <Object>,
            "_files": [
            {
                "filename": <String>,
                "metadata": <Object>,
                "gridFS": <Object>
            }],
            "_clinicalSamples": [
            {
                "sample_name": <String>,
                "metadata": <Object>,
                "gridFS": <Object>
            }]
        }],
    "Form": <Object>,
    "Analysis_results": [
        {
            "name_analysis": <String>,
            "results": <String>
        }]
}
```

As far as this tool is managing sensible data, some security elements are included to reduce the risk of adding these data into the service. Patients' information is secured in the whole database. Advanced Encryption Standard (AES) is used to protect the stored data. This encryption algorithm [4] uses a secret key to encrypt and decrypt the data. The 256-bit keys are used to achieve robust data security. This key is automatically generated during the registration process at the server side.

Additionally, the cryptography keys are stored in a MinIO [4] cluster. This cluster has been deployed in an internal network only accessible from the server where the APIs are allocated, taking also advantage of the security layers provided by MinIO. This reduces the risk of cyber-attacks to the database.

[4] https://min.io/

2.2 Gene Expression Data Mapping

Gene expression data is loaded into FIMED 2.0 by uploading files with the output provided by the different providers. Once the user indicates that an uploaded file corresponds to gene expression data, this data is parsed and translated to a shared internal representation. Thus, it is possible to uniformly manage and analyse different NGS file formats from various providers such as Nanostring, Affymetrix, etc.

The load of such data in FIMED 2.0 ends with a shared structure storing the gene names, class names and gene expression values. In this way, the components analysing such data will be able to translate them to a gene expression matrix. This means, that for each data file we will have the expression levels for each gene included in the specific panel.

FIMED supports RCC (*Reporter Code Count*) format independently of the Nanostring^{TM} panel used. These files provide Code Class, Gene Name, Accession and Count that are directly translated to the shared format. However, this process is not limited to parsing and translation. The translation process includes specific normalization processes that are aware of the essential lane attributes provided by Nanostring^{TM}. At the end of this step, the system will have homogeneous gene expression data as well as other transformations will be conducted to assure high-quality results.

This process can be extended for further versions of the software, adapting the processes for reading the files and normalizing the data to the documentation provided by the providers.

2.3 Gene Expression Pre-processing

Gene expression data pre-processing is mainly focused on the normalization process. Hence, the main objective is to reduce the possible noise that is produced in the gene sequencing. FIMED 2.0 uses a standard normalization process. This process is essential as any noise in the data used for a given analysis will be translated to the analysis results.

RCC files contain quality control flags as positive and negative control genes. Positive control linearity ensures that the samples have a consistent linear relationship. Background correction is conducted with the use of negative control samples. The process done for RCC files is based on the algorithm *geNorm* [5] [12] provided by Nanostring^{TM}.

2.4 Gene Regulatory Network Inference Analysis

Gene regulatory network inference is used to construct gene regulatory networks from previously pre-processed gene expression data. GRNs models are used to represent and predict dependencies between molecular entities [11]. These are composed of genes, in which interactions between genes are represented within a graph model focused on transcription factors (TFs).

[5] https://genorm.cmgg.be/

In this sense, FIMED 2.0 improve the analysis capabilities of FIMED in the context of efficient statistical and machine learning approaches for GRN inference. Originally FIMED integrated two distinguished algorithms (GENIE3 and GRNBoost2) provided in the *arboretum* Python package [6]. However, this version of the tool includes two new gene inference algorithms that will provide the user with a broader comparison of the results in order to improve their analysis capacity. Moreover, FIMED 2.0 provides an ensemble gene regularity inference functionality that enables users examine which algorithms produced similar reconstructions.

GENIE3 is a generic and straightforward algorithm based on feature selection with tree-based ensemble methods. It breaks down the prediction of a regulatory network involving p genes into p separate regression problems. The expression pattern of one of the genes (target gene) is predicted from the expression patterns of all the other genes (input genes) in each regression problem. The importance of an input gene in predicting the expression pattern of a target gene is interpreted as a possible regulatory connection. The network is then recreated by aggregating putative regulatory relationships across all genes to generate a ranking of interactions [8].

GRNBoost2 uses Gradient Boosting Machine regression with early-stopping regularisation to estimate regulatory networks. A tree-based regression model is trained for each gene in the dataset to predict its expression profile using the expression values of a collection of putative transcription factors (TFs). This algorithm is based on the GENIE3 architecture [13].

PANDA (Passing Attributes between Networks for Data Assimilation) is a message-passing model that integrates protein-protein interaction, gene expression, and sequence motif data to reconstruct genome-wide, condition-specific regulatory networks as a model. In this regard, the networks that are generated are more accurate than those constructed using individual data sets. Gene regulatory network generating with PANDA can also capture information about specific biological mechanisms and pathways that other methodologies had ignored [5].

LIONESS is a linear framework to relate a set of networks, each representing a different biological sample. The average of individual component networks reflecting the contributions of each member in the input sample set can be thought of as an "aggregate" network predicted from a collection of N samples [10].

Gene Regulatory Network Ensemble For further analysis, an ensemble approach has been developed in this proposal as a gene regulatory network inference made from the four prior networks (GENIE3, GRNBoost2, PANDA, LIONESS).

[6] https://arboretum.readthedocs.io/en/latest/

The ensemble approach has been designed since network inference algorithms are naturally noisy, it remains a challenge to identify whether these changes represent real cellular responses or whether they emerged by random coincidence. In this sense, the ensemble internally develops a voting system in order to rank the top-k edges compose by the similar outputs of a set of GRNs, thus users can examine the top-k edges produced by similar reconstructions of the GRN algorithms.

2.5 Gene Regulatory Network Inference Visualization

One of the advantages of FIMED 2.0 is its power related to visualization features thanks to the availability of better graphic plotting, where users can interactively explore the constructed network. Many interactive visualizations allow users to actively move the network and examine the connections between nodes, allowing users to see the network's structure in detail. In this sense, users can choose from different network representation layouts: Circular layout or Force-directed layout, as shown in Figure 2. It is worth noting that FIMED 2.0 offers a new graph visualisation in which clicking on a given gene will highlight this gene and its related neighbours and the information associated with them (Figure 2C).

These rich set of visualization tools allow users to observe the most important nodes representing genes and arcs representing interactions between them. By changing the arc form or length, the arcs can express the strength of the interaction. Users can compare different patient samples or patients at the same stage of sickness in this way.

Additionally, as mentioned before, an ensemble powerful visualization tool has been developed combining various GRNs models, where users can examine the top-k edges between gene interactions. Each similar reconstructions are represented in different edges colors. In this way, users can observe edges frequency in different colours to see the most important gene interaction in the ensemble voting system as depicted in Figure 2D, where grey edge colour represent frequency of 1, blue edge frequency of 2, green edge colour frequency of 3 and red edge colour frequency of 4.

3 Use Cases

With the aim of showing the potential of using the new functionalities of FIMED 2.0, the tool has been tested with real-world scenarios involving patients with advanced melanoma [14], as in the previous version of FIMED. Thus, we have validated the new analytical functionalities and visualisation techniques producing appropriate analysis and visualisation in cancer research. For this proposal, we have used the FIMED 2.0 online interface to enter the clinical information of two Melanoma patients. In this sense, for this clinical trial, we designed a customised Electronic Case Report Form (eCRF). Then, the clinical information of the patients was entered into the tool. As seen in Code Snippet 1.2, this clinical case comprises six simple fields and one composed field. Thanks

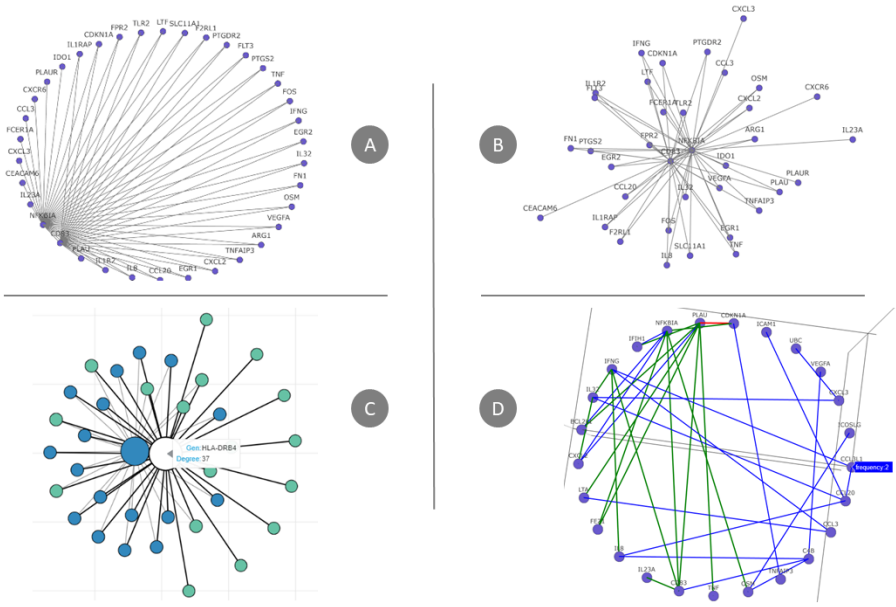

Fig. 2: Gene regulatory network representations with different layouts: Circular layout or Force-directed, and dynamic plotting

to MongoDB's flexibility, the primary database structure can be increased in a customised manner.

In addition, other files providing gene expression assays related to the patient have been loaded in FIMED2.0. As a result, new meta-data fields in the gene expression files have been introduced to offer more information to the samples.

Gene expression data from Melanoma cancer have been taken from a specific panel (Immune Profiling Panel Nanostring^{TM}) in RCC format (under MIT License). These files have been normalised through a housekeeping method. Users can find 12 RCC files with the gene counts stored in FIMED 2.0 (*Help section*: *Supplementary material*). Each file has associated with the experiment date and the patient's code.

FIMED 2.0 has been deployed on our servers to enable users to explore the new functionalities, where users can manage their patient data or test ittest it using sample data given by the demo user provided [7]. This demo user includes anonymized patient data to allow new users to see an example of how their databases may be built. Users can also establish a new free account in which each user will have an independent workspace to design a particular database schema for their clinical trial.

Code Snippet 1.2: Data Schema in Melanoma use cases

[7] Demo user grants: username *"iwbbio"* and password *"demo"*

```
{   "Form":
    {
        "_id":<ObjectId()>,
        "Attributes":
        {
            "Patient Code": <Number>,
            "Sex": <String>,
            "Birth Date": <Date>,
            "Treatment response": <String>,
            "Observations": <String>,
            "Diabetes": <Checkbox>,
            "Hospital admission":
            {
                "Hospital name": <String>,
                "Hospital address": <String>
            }
        }
    }
}
```

3.1 Use case: Gene Regulatory Network

In this section, the new functionalities in FIMED 2.0 are shown from the GRNs point of view. As exposed above, new GRNs algorithms have been added to the tool, as well as new features in the visualization part that improves the ability of users to discover important gene-to-gene interactions and to inspect the topology of the network thanks to the availability of better graphic plotting.

In Figure 3 can be observed the selection panel of FIMED 2.0 that allow users to perform gene regulatory network analysis and visualizations. Five GRNs visualizations have been performed, corresponding to each of the GRNs algorithms provided in FIMED 2.0. An experiment has been carried out which consist in inferring a set of different networks (GENIE3, GRNBoost2, PANDA, LIONESS) and compare the results of each of the networks. Besides, the ensemble algorithm based in a voting system has been executed in order to provide users an entirely deep insight of the most important gene interactions coming from the similar reconstructions of the GRNs algorithms.

Users will then be able to distinguish the frequency that each interactions between genes are repeated through similar GRNs outputs since each frequency is represented with a different edge colour (frequency 1: grey, frequency 2: blue, frequency 3: green, frequency 4: red). Therefore, the most important gene interactions (highest frequency) are represented with a red edge colour.

Furthermore, only the most variable gene expression levels as a fraction of the total number of genes in the panel can be extracted using a sliding parameter. A statistical cutoff parameter is also supplied to limit the maximum number of linkages in the network, which improves visualisation because it focuses only on the most relevant genes and their interactions. It's worth mentioning that these

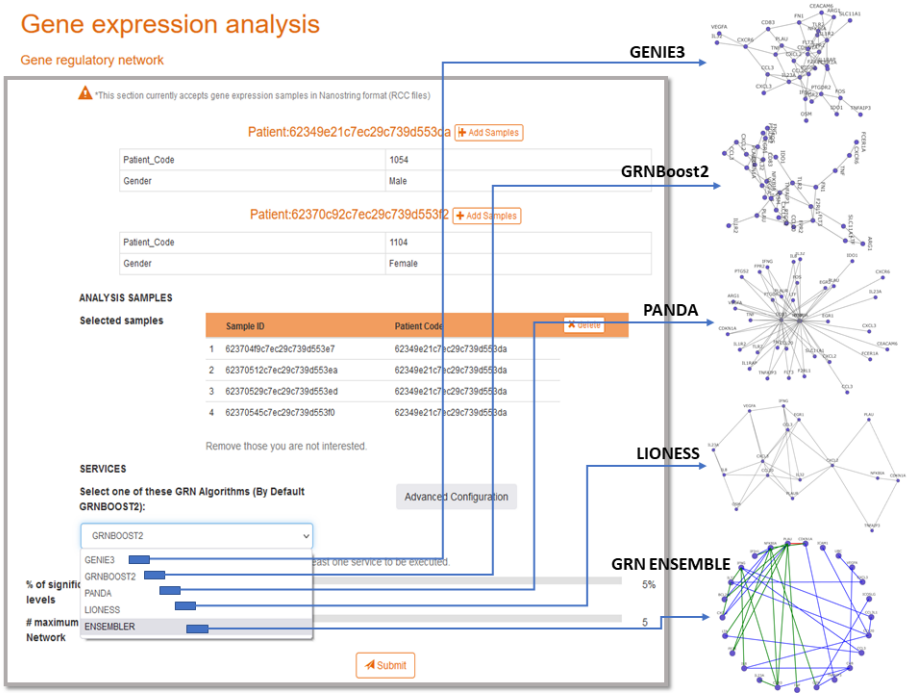

Fig. 3: Selection panel of FIMED 2.0 that allow users to perform gene regulatory network analysis and visualizations from gene expression data

features might provide clinicians with new information for improving treatment outcomes by allowing users to find genes and gene interactions that could be utilised as diagnostic and prognostic indicators and focused therapy.

4 Conclusion

In this paper, we present FIMED 2.0, a software tool for flexible clinical data collection, management, analysis and visualization of gene expression data in the practice of clinical assays in different studied diseases. It is released freely available on the web for the community at https://khaos.uma.es/fimedV2/.

The current version improves the previous version of FIMED in terms of Gene Regulatory analysis inference. New distinguished GRNs algorithms (PANDA and LIONESS) have been integrated in the tool that will provide users with better analysis capacities in order to increase their GRNs understanding. Besides, an ensemble has been designed based on a voting system of the similar reconstructions network from a set of four GRNs algorithms (GENIE3, GRNBoost2, PANDA and LIONESS). Moreover, new visualization features have been added guaranteeing users new ways of exploring gene networks to clearly inspect the topology of the network, thanks to the availability of better graphic plotting.

All these new functionalities of the tool have been tested in a use case conducted with real-world gene expression data from Melanoma cancer. These data have been stored in FIMED 2.0 so that users can explore the tool with a demo user *"iwbbio"* and password *"demo"*.

Specific lines of future work include incorporating the compatibility with additional use cases. Thus, integrating adaptability for more gene expression file formats, other clinical studies on different diseases, and other algorithms such as clinical image analysis.

Acknowledgement

This work has been partially funded by the Spanish Ministry of Science and Innovation via Grant TIN2017-86049-R (AEI/FEDER, UE) and Andalusian PAIDI program with grant P18-RT-2799.

References

1. Brandt, C., Deshpande, A.M., Lu, C., Ananth, G., Sun, K., Gadagkar, R., Morse, R., Rodriguez, C., Miller, P.L., Nadkarni, P.M.: Trialdb: A web-based clinical study data management system. In: AMIA Annual Symposium Proceedings. vol. 2003, pp. 794–794. American Medical Informatics Association (2003)
2. Cavelaars, M., Rousseau, J., Parlayan, C., de Ridder, S., Verburg, A., Ross, R., Visser, G.R., Rotte, A., Azevedo, R., Boiten, J.W., et al.: Openclinica. In: Journal of clinical bioinformatics. vol. 5, p. S2. Springer (2015)
3. Cramon, P., Rasmussen, Å.K., Bonnema, S.J., Bjorner, J.B., Feldt-Rasmussen, U., Groenvold, M., Hegedüs, L., Watt, T.: Development and implementation of progmatic: A clinical trial management system for pragmatic multi-centre trials, optimised for electronic data capture and patient-reported outcomes. Clinical Trials **11**(3), 344–354 (2014)
4. Daemen, J., Rijmen, V.: The design of Rijndael: AES-the advanced encryption standard. Springer Science & Business Media, Belgium (2013)
5. Glass, K., Huttenhower, C., Quackenbush, J., Yuan, G.C.: Passing messages between biological networks to refine predicted interactions. PloS one **8**(5), e64832 (2013)
6. Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J.G.: Research electronic data capture (redcap)—a metadata-driven methodology and workflow process for providing translational research informatics support. Journal of biomedical informatics **42**(2), 377–381 (2009)
7. Hurtado, S., García-Nieto, J., Navas-Delgado, I., Aldana-Montes, J.F.: Fimed: Flexible management of biomedical data. Computer Methods and Programs in Biomedicine **212**, 106496 (2021)
8. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., Geurts, P.: Inferring regulatory networks from expression data using tree-based methods. PloS one **5**(9), e12776 (2010)
9. KRENN, R.: Design and development of a web-based clinical trial management system

10. Kuijjer, M.L., Tung, M.G., Yuan, G., Quackenbush, J., Glass, K.: Estimating sample-specific regulatory networks. iScience **14**, 226–240 (2019). https://doi.org/https://doi.org/10.1016/j.isci.2019.03.021, https://www.sciencedirect.com/science/article/pii/S2589004219300872
11. McCall, M.N.: Estimation of gene regulatory networks. Postdoc journal: a journal of postdoctoral research and postdoctoral affairs **1**(1), 60 (2013)
12. Mestdagh, P., Van Vlierberghe, P., De Weer, A., Muth, D., Westermann, F., Speleman, F., Vandesompele, J.: A novel and universal method for microrna rt-qpcr data normalization. Genome biology **10**(6), R64 (2009)
13. Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., Aerts, S.: Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics **35**(12), 2159–2161 (2018)
14. Navas-Delgado, I., García-Nieto, J., López-Camacho, E., Rybinski, M., Lavado, R., Guerrero, M.Á.B., Aldana-Montes, J.F.: Vigla-m: visual gene expression data analytics. BMC bioinformatics **20**(4), 150 (2019)
15. Nguyen, L., Shah, A., Harker, M., Martins, H., McCready, M., Menezes, A., Jacobs, D.O., Pietrobon, R.: Dados-prospective: an open source application for web-based prospective data collection. Source code for biology and medicine **1**(1), 7 (2006)
16. Ou, M., Ma, R., Cheung, J., Lo, K., Yee, P., Luo, T., Chan, T., Au, C.H., Kwong, A., Luo, R., et al.: Database. bio: a web application for interpreting human variations. Bioinformatics **31**(24), 4035–4037 (2015)
17. Pereira, R., Oliveira, J., Sousa, M.: Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. Journal of Clinical Medicine **9**(1) (2020). https://doi.org/10.3390/jcm9010132, https://www.mdpi.com/2077-0383/9/1/132