VarCopy: a Visual Exploratory Data Analysis Platform for Copy Number Variation Studies

Fabio Bove, Federica Mandreoli,
Riccardo Martoglia, Valentino Pisi
FIM - University of Modena and Reggio Emilia, Italy
fabio.bove.dr@gmail.com, federica.mandreoli@unimore.it,
riccardo.martoglia@unimore.it, valentino.pisi@gmail.com

Cristian Taccioli, Chiara Vischioni
MAPS - University of Padova, Italy
cristian.taccioli@unipd.it, chiara.vischioni@phd.unipd.it

***Abstract*—The study of such a complex phenomenon as cancer, which depends on several but unexplored and unclear factors, needs new ways to visualize, analyze and combine different data both on species characteristics and genes function. To this respect, we propose a novel platform, named VarCopy, supporting visual Exploratory Data Analysis (EDA) in the context of Copy Number Variation (CNV) data. The platform will be publicly available as a web application soon, and is, to our best knowledge, the first tool allowing visual, interactive exploration and analysis of the CNV landscape of multiple species, allowing the identification of new target genes that might be useful for biomedical research.**

***Index Terms*—Copy Number Variations, interactive visualization, Exploratory Data Analysis, scalable data science, data analysis models**

I. Introduction

Recently, biomedical research has looked at the study of those species having peculiar properties in terms of cancer resistance and high longevity rates, producing fundamental discoveries on the mechanisms that might protect an organism from the development of tumorigenesis [1].

A promising research is the study of *Copy Number Variations* (CNVs), that are defined as the number of gene copies and might be related to cancer initiation and/or genome instability [2].

In this context, we developed VarCopy, a new tool able to compare and correlate the CNVs landscape of different organisms, allowing the identification of genomic region of high instability. The combination of different IT technologies and strategies, especially *Exploratory Data Analysis* (EDA) [3] tools, gives VarCopy the possibility to offer visual and interactive ways to explore, summarize, and analyze the large amount of data in an easy, user-friendly and fast manner.

The contributions of the VarCopy platform, whose work is still in progress and which will be made publicly available as a web application soon, are the following:

- the platform is based on a unique of its kind database that combines CNV data among different species with other vital parameters;
- the information is retrieved from public on-line genomic libraries;
- exploratory data analysis is made possible on the large amount of information contained in its database through a variety of tools, allowing researchers to: (a) freely combine and use different *Data Analysis Models* (DAMs) to generate visual and interactive reports and plots, (b) access additional related data extracted from reference on-line sources, (c) perform advanced personalized searches by means of custom queries allowing an infinite number of search possibilities on the data;
- the client technologies are coupled with server-side optimizations inspired by scalable data science [4] (including parallel threads and dynamic results caching) that together enable the needed real-time interaction experience on the large amount of data.

In particular, the use of DAMs, together with statistical measurements, will enable researchers to easily identify new patterns, highlighting which are the genes linked to cancer resistant species or long-living organisms.

The paper is structured in the following way: Section II discusses related works, Section III is devoted to the description of the database underlying the platform, whereas the platform itself is described in Sections IV (high-level overview), V (EDA functionalities) and VI (implementation and performance evaluation). Finally, Section VII concludes the paper and discusses future works.

II. Related Works

EDA is a well-established statistical tradition that provides conceptual and computational tools for discovering patterns in a data science context [3]. Indeed, researchers often struggle to develop hypotheses despite the abundance of data available to them. In recent years, EDA and data visualization techniques [5] have been suggested by data scientists [6], [7] as an effective step for pattern and hypothesis generation in a data science process. Moreover, scalable data science is also empowering the domain sciences, healthcare, humanities, governance, journalism, and other fields to study phenomena at scales and granularities never before possible [4], for instance in the field of cloud data mining [8] and data streams visualization [9].

The Ensembl CNV tool [10] is the world's largest source of information on the functions of genes from different species and includes human-readable and machine-readable information about this subject. The Ensembl section devoted to CNVs provides a searchable and integrated database that can be only used through APIs. While this tools is certainly very powerful in the field, it also shares a number of shortcomings:

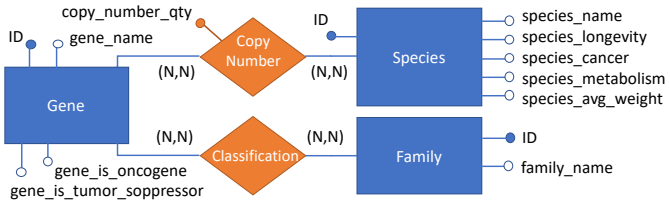

Fig. 1. VarCopy Database structure: ER Diagram

- the search speed is typically slow and quite time-expensive for the user;
- the "heavy" graphic style can be often difficult to read, thus making interaction less friendly and immediate;
- above all, they lack proper data visualization techniques.

Our platform is designed to overcome the limitations on the above aspects. In particular, the client offers ad-hoc visualization facilities, with convenient plots to represent the data and their distributions, while the scalable data science implementation techniques include ad-hoc optimizations for a faster and less expensive information management and real-time results and interaction. Going even further, the EDA tools enable recursive searches directly from the output plots and visualizations, and also personalized searches including an advanced search where users can specify through a custom query complex search patterns on the database.

III. The Dataset

VarCopy platform allows the user to access a unique dataset that combines CNV data with other information related to their cancer rates, longevity and vital parameters. Data are extracted from Ensembl [10] and NCBI [11].

Figure 1 shows the structure of VarCopy database that is made up of five tables. The total amount of data for each table is the following:

- 21,036 tuples for the Gene table;
- 9,996 tuples for the Family table;
- 192 tuples for the Species table;
- 3,005,109 tuples for the Copynumber table;
- 835,088 tuples for the Classification table.

It is worth noting that in the Gene table there are two boolean attributes, gene_is_oncogene and gene_is_tumor_suppressor, that record whether the instance is known to be an oncogene or a tumor suppressor, respectively. This information is available only for a small portion of the available genes, i.e. 11.20% of the stored genes. In the Species table, instead, the boolean attribute species_cancer distinguishes species in cancer prone and cancer resistant ones, while species_longevity assesses whether the instance is a long-lived species. This information is currently available for 13% of the species.

IV. Platform Overview

VarCopy platform aims at supporting the identification and hypothesis testing of potential tumor suppressor or oncogenes by implementing advanced methods that enable the efficient and effective visualization and analysis of the large amount of information contained in its database. Following the trend of recent data engineering approaches [3], [4], [6], the platform implements EDA functionalities by allowing users to issue different kinds of data analysis requests and by showing results through interactive plots users can browse and select to issue further requests. These graphical representations can be *Descriptive Analysis Models (DAMs)* that highlight CNV trends of groups of species known to have particular properties, or contained in *target reports*, showing CNV statistical summaries of the target genes or species.

VarCopy is implemented as a Web Application, making all the analysis and visualization tools easily accessible to researchers. The Web Application is organized in four areas and made up by several components as shown in Figure 2 (areas shown in orange, components in blue):

- The "**Home**" area welcomes researchers and proposes basic overview explanation and plots; moreover, it allows basic search functionalities on genes, species and families;
- The "**Species Exploration**" and "**Genes Exploration**" areas provide advanced search functionalities on species and genes, respectively;
- The "**Analysis**" area proposes several DAMs and a personalized query form for ad-hoc searching and filtering;
- Target reports implement different models, measures and statistical tests over target genes or species and allow users to interact with them.

Following the "exploratory" philosophy, where the EDA process is flexible and the result is uncertain [3], different kinds of interactions are enabled so that users can incrementally build their own data exploration path according to their deductions (see Section V for a detailed description of the EDA functionalities). For instance, a researcher could start his/her analysis by issuing an advanced search request that filters on "LOXODONTA AFRICANA", a species that has a low rate of cancer incidence. The user will be shown a target report that contains an interactive plot about the number of copies of all the genes in its genome. The user can be attracted from a gene having a high number of copies in the selected species, and thus can study its behaviour. (S)he will then obtain a target report on the selected gene. The report contains a box plot showing that this gene has a high mean value of copies in cancer resistant species, whereas it is very low in cancer prone ones. The researcher can then decide to continue the investigation in vitro.

In order to efficiently support the above EDA functionalities, VarCopy implements data caching and parallel threading solutions enabling rapid and effective response from the platform, also against expensive requests. The design of such mechanisms is inspired by the "Scalable Data Science/Analysis" trend [4], which is a current topic in research [8], [12], and will be described in Section VI.

V. Exploratory Data Analysis Functionalities

EDA functionalities constitute the core of the platform analytical power. Data analytics requests can be:

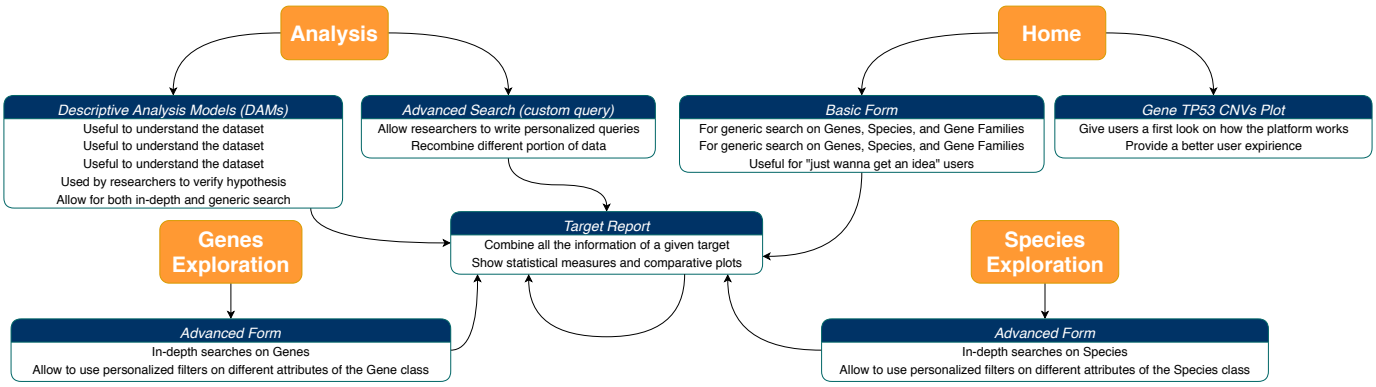

Fig. 2. Platform schema: Interactions between areas of the platform (shown in orange) and their components (in blue)

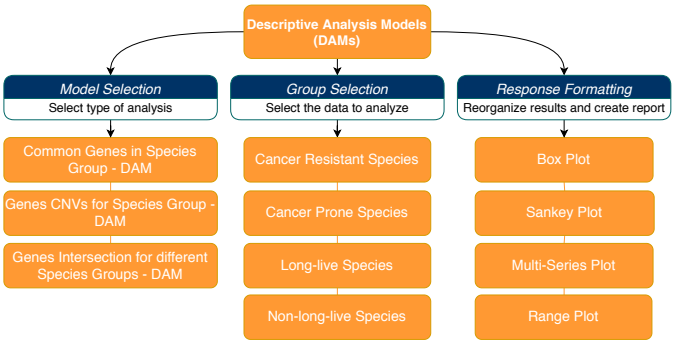

Fig. 3. Descriptive Analysis Models (DAMs) schema: a *model* is applied to a *group* and a resulting report is created as *response*

- **triggered** in different ways: (a) by exploiting the *forms* associated with each DAM; or (b) by *interacting*, in a recursive way, with one of the plots returned from previous analyses (i.e., from a DAM or target report);
- **targeted** on different objects: (a) specific database *entities* (i.e., genes, species, families and groups) as prescribed by each DAM; or (b) *custom queries* for even greater flexibility.

A. Descriptive Analysis Models

One essential tool for researchers studying CNVs are Descriptive Analysis Models, supporting effective exploratory data analysis over CNV properties related to both genes and species. VarCopy platform provides a choice of different DAMs for different analytical needs. DAMs are dynamically generated by selecting one of the available *models* and by applying it to a specific target *group* (see Figure 3). Four default groups are provided: cancer resistant and cancer prone species, long-living and non-long-living species. The analysis output is generated through *response formatting*, and the results are sorted on the basis of different *ranking* strategies.

Three different types of Descriptive Analysis Models have been devised, each providing specific tools and interfaces (see Figure 4 for an overview of their UI):

- **Common Genes Model**: it provides a visual overview of the CNVs of genes within the species of a group (e.g., cancer prone or long-living species). The order of the results for each series is for decreasing CNV;
- **Genes Copy Number Quantity Model**: useful to know which are the genes with the highest CNV in the species of a group and how their statistical measures fluctuate. This model allows a fast visualization of statistical values such as mean, standard deviation, max and min. Results are sorted by decreasing maximum number of gene copies;
- **Genes Intersection Model**: the differences in terms of copies of a gene for the species of two different groups could highlight if a gene is interesting for the analysis, and if it can be correlate with tumorigenesis. Furthermore, it shows the maximum number of copies in one of the species within the two groups, so that researchers can identify species to study more in detail.

Figure 5 shows an example of interaction with the common genes DAM. This DAM contains a plot showing the distribution of copies for each gene across all the species belonging to the group, selected via the drop-down menu (upper left of figure). Users can filter the samples and obtain more specific plots, showing the data concerning only the selected species (bottom right of figure). An on-mouse-over interaction has also been implemented on the plot results to show the label concerning the highlighted target. Moreover, via the button on the upper right of the plot, users can download the currently visualized model in a number of formats. Finally, by interacting with the results shown in the plot or via the text-box in the upper right part of the UI, researchers can obtain reports based on the chosen target.

B. Target reports

Target report provides detailed information about selected genes or species. Reports contain plots built from the data in the dataset that relates with the target and links to the external knowledge sources, i.e. Gene Cards [13], Gene Ontology (GO) [14] and ADW [15]. For instance, additional resources provided for target genes include families and function descriptions, mean copies of the genes and variance in cancer prone and cancer resistant species with p-value and t-test statistical

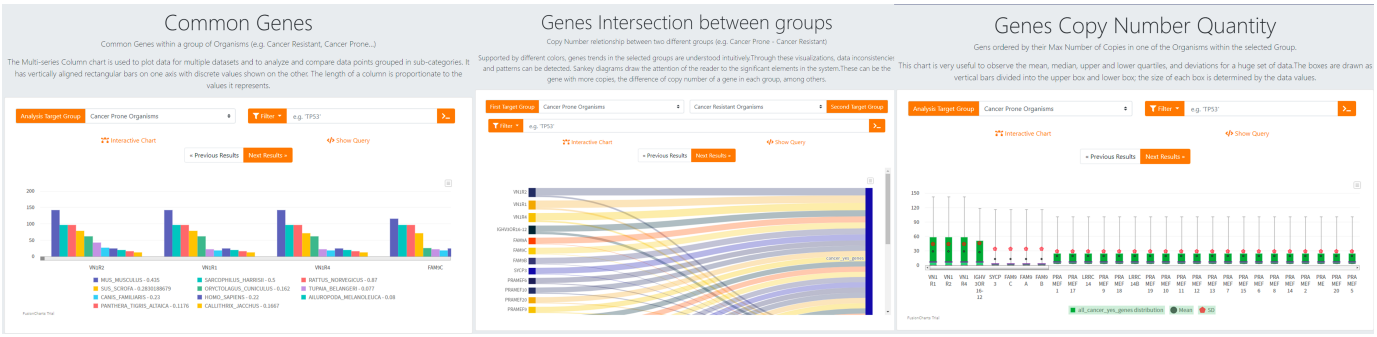

Fig. 4. User interfaces of the three main DAMs: Common genes (left), Genes Intersection (center) and Genes Copy Number Quantity (right)

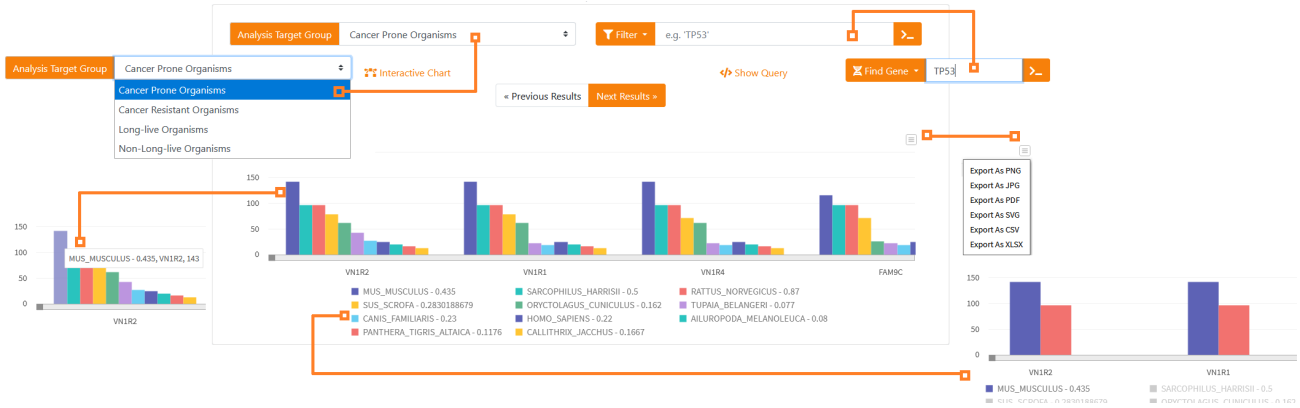

Fig. 5. Descriptive Analysis Models (DAMs) UI - Interaction example with common genes model

measurements, while for species the platform returns cancer mortality, longevity, metabolism and average weight and number of copies of oncogenes and tumor suppressors within the chosen sample genome. P-value and t-test are obtained with the "two tailed" Welch T-Test for unequal group of samples with different variance [16]. Their values are essential to find new possible oncogenes or tumor suppressors: when a gene has a p-value lower or equal to 0.05, researchers can reject the null hypothesis, this means that the gene discriminates well the groups and its high CNVs could have some relationship with low cancer rates.

C. Custom queries

As far as custom queries are concerned, they can be specified in a dedicated free text form (see Figure 6). Unlike other forms, where the queries have limited combinations, and the possible values are suggested by autocomplete mechanisms, custom queries give the user infinite search possibilities by means of an easy to understand syntax. The syntax allows the user to request any column of any table in the database (**`show`** clause) and to filter the retrieved information (**`filter`** and **`exclude`** clauses). For instance, the following query retrieves the genes contained in the Homo Sapiens species:

```
show:(specie_name, gene_name, copy_number_qta)
filter:(specie_name=HOMO SAPIENS)
```

D. Ranking strategies

Data analytics requests often output a very large number of results, either genes or species. It is therefore of utmost importance to implement ranking strategies that reward the genes or species that are the most representatives in the selected group. For the majority of the plots, the ranking criterion relies on the number of copies. Moreover, VarCopy introduces a new statistical ranking mechanism, named "genes trend", which privileges genes showing an interesting behavior:

- an elevated number of copies for cancer-resistant species;
- a low distribution of the number of copies for the rest of the population.

The ranking formula for the gene g is the following:

$$rank(g) = \frac{max(Copies^{g}_{CRSpecies}) - avg(Copies^{g}_{AllSpecies})}{max(Copies^{g}_{CPSpecies})}$$

where "CRSpecies" and "CPSpecies" stand for cancer resistant and cancer prone species, respectively.

VI. Implementation strategies and performance

In order to provide the powerful and user-friendly features discussed in the previous sections with efficient performances, VarCopy has been implemented with advanced server- and client-side technologies and optimizations, also inspired by scalable data science [4], that together enable the needed real-time interaction experience on the large amount of data.

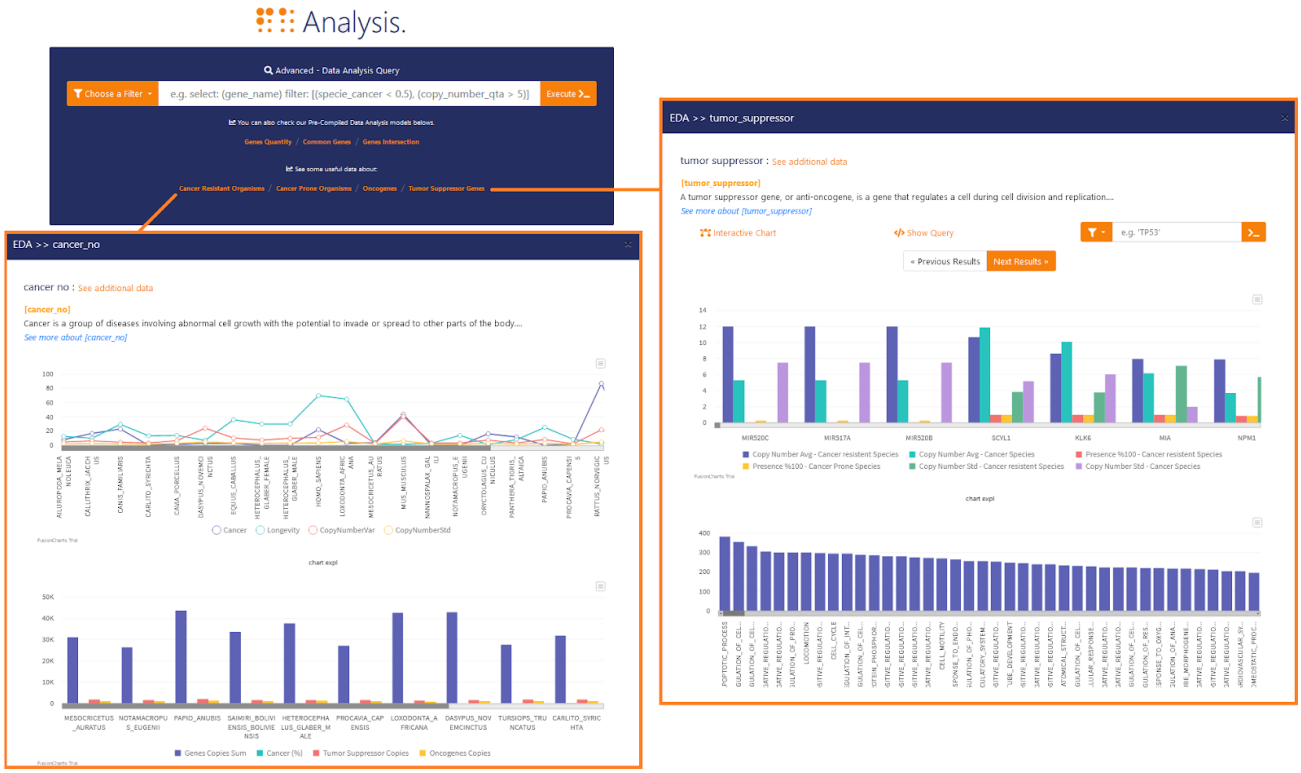

Fig. 6. Custom Query and additional data panels

Sections VI-A and VI-B give an overview of the exploited technologies and optimizations; Section VI-C concludes the discussion with performance evaluation.

A. Exploited Technologies

The platform is written in Python and is based on some of the latest technologies for efficient client-server web application development and data management: Django[1], FusionCharts[2], PostgreSQL[3] and Pandas[4] for server-side data management, and Scikit-Learn[5]. Django allowed us to create web application, integrating administrator interface for data manipulation and managing the web application client and server communication. FusionChart enabled client plot creation and drawing. PostgresSQL was chosen for database storage because of its balanced performances, and for its advantage to efficiently write and manage large amounts of data on disk. Pandas was used for data caching, manipulation and analysis, providing the facilities for the management of dataframes and time series. Finally, the Scikit-Learn statistical library was chosen to support the computation of the statistical values (p-value and t-test) presented in the reports.

[1]http://www.djangoproject.com/
[2]http://www.fusioncharts.com/
[3]http://www.postgresql.org/
[4]http://pandas.pydata.org/
[5]http://scikit-learn.org/

B. Server- and client-side Optimizations

The power of the advanced searches and interactive functions offered by the platform would be nearly useless without very efficient processing strategies allowing immediate response time even for complex requests. In particular, the platform forms require the querying, retrieval and presentation of large amounts of data. For efficient processing, VarCopy implementation exploits server-side caching and threading as well as client-side caching and chunking optimizations.

Server-side caching. The code of each query is dynamically composed from the DAMs forms. The choice is to have precompiled subqueries stored according to the model and the selected groups. Using the concept of "saved and quick to read data" and through Pandas dataframes files, we want to create highly modular "caches" partitioned according to the contained information. Each of these files contains the result of a query: in particular, we cache information about copy numbers with relative gene names, families joined with their classifications, cancer-prone and cancer resistant species names and relative common genes, species. These structures help us as they are already indexed and enable many types of different joins with structures of the same kind. In this way, the response latency, also for custom queries, is considerably reduced by avoiding the recalculation of common queries.

Server-side threading. Further response latency reduction is achieved by threading, enabling the parallelization of the processes needed for the construction of the form responses. Once query results are received, the platform starts two

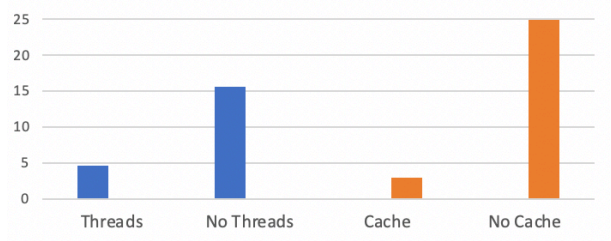

Fig. 7. Response time tests (time in seconds, mean of 20 executions)

different threads. The first thread deals with the construction of the structures necessary for formatting report table data and retrieving additional resources from external sites. The second thread is dedicated to the construction of the plots. Moreover, in the case of advanced searches (custom queries), parallelism is even higher: the query is parsed in order to identify the contained fields, then a thread that will retrieve cached data is created for each one of them. All the threads' retrieved data is finally merged and filtered.

Client-side caching and chunking. User requests inevitably produce a high number of results, typically close to 20,000 records and beyond. Trying to visualize this data without any pre-processing would lead to an inefficient plot rendering, resulting in a bad user experience and compromising the platform functionality. The solution to this problem has been the division of the dataset (result of each DAM) in multiple chunks (this process is done client-side) and the exploiting of client-side result caching. While the data of the first chunk is directly visualized, the other cached results can be subsequently used without the need to send new requests to the server.

C. Performance evaluation

Numerous performance evaluations were performed on the platform. Due to lack of space, in this section we will focus on the tests we performed to quantify the response time benefits offered by server-side caching and threading. All test were performed on a server with AMD Ryzen 1920X CPU, 32GB RAM, 256GB SSD and 2TB HDDs.

VII. Conclusions and Future Works

In this paper we presented VarCopy, a platform supporting visual EDA in the context of Copy Number Variation (CNV) studies. The platform is aimed to assist researchers in exploring the relationships among the information that has been collected so far on the subject. It also sets itself the goal of growing together with the research sector to which it is aimed, trying to become a tool on which researchers can rely.

Figure 7 shows two response time comparisons. The first one (left side, blue bars) is relative to the mean response time measured for standard interaction / querying on the genes and species exploration forms, with and without threading. As we can see, threading optimizations almost produce a 4x speedup, making all interactions almost real-time. The second comparison (on the right side, orange bars) quantifies the advantages offered by server-side caching, in particular in the complex case of custom queries. A large number of queries are avoided on the database, leading to a sharp reduction in response time. In particular, the required time depends on the number of tables involved in the user query. For this test, we considered the worst case (all tables involved): the measured speedup reaches 10x, going from 25 seconds (no caching) to less than 3 seconds (caching active).

Thanks to this platform, researchers have already identified important genes related to carcinogenesis. These include genes that have never been studied in depth but that, through the analysis on our platform, have shown important characteristics for groups of species with low cancer mortality rates.

The platform will be made publicly available through web access, as done in the past for other genomic exploration tools [17]. In order to give even more answers to the currently unknown causes of tumors onset we plan, in the future, to expand VarCopy capabilities, for instance by incorporating new algorithms for the classification of genes and the discovery of such parameters that can cooperate in the cancer onset. As far as we know, VarCopy is the first tool allowing the researchers to compare CNVs from different species and identify those genes that appear to lead to a genome instability.

References

[1] M. Serrano, "Unraveling the links between cancer and aging," *Carcinogenesis*, vol. 37, no. 2, p. 107, 2015.

[2] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 85–97, 2006.

[3] J. T. Behrens, "Principles and procedures of exploratory data analysis," *Psychological Methods*, vol. 2, no. 2, pp. 131 – 160, 1997.

[4] "ACM Sigmod Blog: Scalable Data Science: a new Research Track Category at Pvldb vol 14 / Vldb 2021," http://wp.sigmod.org/?p=3033.

[5] E. R. Tufte, *The Visual Display of Quantitative Information*, 2nd ed. Graphics Press, 2001.

[6] R. Schutt and C. O'Neil, *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media, Inc., 2013.

[7] X. Ma, D. Hummer, J. Golden *et al.*, "Using visual exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research," *ISPRS International Journal of Geo-Information*, vol. 6, no. 11, p. 368, Nov 2017. [Online]. Available: http://dx.doi.org/10.3390/ijgi6110368

[8] D. Talia, "A view of programming scalable data analysis: from clouds to exascale," *Journal of Cloud Computing*, vol. 8, 12 2019.

[9] B. Breve, L. Caruccio, S. Cirillo, V. Deufemia, and G. Polese, "Visualizing dependencies during incremental discovery processes," in *Proc. of EDBT/ICDT Workshops*, ser. CEUR Workshop Proceedings, vol. 2578, 2020.

[10] "Ensembl Genome Browser," http://www.ensembl.org/.

[11] "NCBI Genome," http://www.ncbi.nlm.nih.gov/genome/.

[12] W. Tang, J. Wilkening, N. Desai, W. Gerlach, A. Wilke, and F. Meyer, "A scalable data analysis platform for metagenomics," in *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, 10 2013.

[13] "GeneCards: The Human Gene Database," http://www.genecards.org/.

[14] "The Gene Ontology Resource," http://geneontology.org/.

[15] "ADW - Animal Diversity Web," http://animaldiversity.org/.

[16] G. Ruxton, "The unequal variance t-test is an underused alternative to student's t-test and the mann-whitney u test," *Behavioral Ecology*, vol. 17, 04 2006.

[17] V. Lomonaco, R. Martoglia, F. Mandreoli, L. Anderlucci, W. Emmett, S. Bicciato, and C. Taccioli, "Ucbase 2.0: Ultraconserved sequences database (2014 update)," *Database: the journal of biological databases and curation*, vol. 2014, 2014.