snpStorage: a database to store and share experimental SNPs data sets

Giuseppe Agapito[1,3], Mario Cannataro[2,3]

[1]Dep. of Legal, Economic and Social Sciences, University Magna Græcia,Catanzaro, 88100, Italy
[2]Dep. of Medical and Surgical Sciences, University Magna Græcia, Catanzaro, 88100, Italy
[3]Data Analytics Research Center, University Magna Græcia, Catanzaro, 88100, Italy

Abstract

The last years have seen a continuous improvement in sequencing and genotyping technologies. These improvements have spawned a rapid advancement in bioinformatics databases and software relating to the collection and analysis of genetic data. Genetic data holds several types of variations such as: *indels*, *microsatellites*, *copy number variants (CNV)*, and *Single Nucleotide Polymorphisms (SNPs)*. SNPs are the most abundant type of genetic variation and are now the principals raw material underlying most genetic studies and databases. SNPs can be used as potential biomarkers to investigate cancer growth or progression, and planning more effective treatment regimens. To make data quickly accessible to the users, e.g., to face the burden related to the relation of several experiments and to allow an easy and controlled sharing of genotyping data among different labs, the need for specialized SNP databases arises. For these reasons, we developed *snpStorage* (and related querying service), a database to collect experimental SNPs data integrated with clinical data if available. *snpStorage* collects experimental SNPs data in pharmacogenomics studies, obtained by using DMET (*Drug Metabolism Enzymes and Transporters*) platform and subsequently enriched with clinical data. The main contributions of snpStorage are *i)* automatic translation, normalization and integration of SNP data sets with clinical data and mapping in relational schema. *ii)* A repository where research laboratories can store data (anonymously), merge and handle SNP data sets for more specific analysis; *iii)* To provide medical professionals a software instrument to facilitate clinical decisions based on the individual's genome, and tailoring health care services to the patient's genotype.

Keywords

Databases, Single Nucleotide Polymorphisms (SNPs), SQL, NoSQL, Genomics, Integrative Bioinformatics

1. Introduction

The last years have seen exponential growth in the investigation of genetic and genomic variations as more genomes have been sequenced, corresponding with the continuous improvement in sequencing and genotyping technologies.

These improvements have spawned a rapidly advanced in bioinformatics databases and softwares relating to the collection, management, and analysis of genomics data.

Genomics data include the variations that are very common and useful in genomics studies,

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy
agapito@unicz.it (G. Agapito); cannataro@unicz.it (M. Cannataro)
https://kmitd.github.io/ilaria/ (M. Cannataro)
0000-0002-0877-7063 (G. Agapito); 0000-0001-7116-9338 (M. Cannataro)

including *indels*, *microsatellites*, *copy number variants (CNV)*, and Single Nucleotide Polymorphisms (*SNPs*). SNPs are the most abundant type of genetic variations, and are now the principal raw material underlying most genetic studies. SNPs are common mutations in the population; the most common and frequent SNP type is the substitution of a single DNA's bases. The DNA's bases are Adenine, Cytosine, Guanine, and Thymine (while Thymine is replaced from the Uracil into the RNA) in short the nucleotides are $(A, C, G, \{T/U\})$. Among the other types of variations SNPs are mainly the easiest to observe and the most useful and widely applied markers in genetic and epigenetic studies. The continue development and improvement of the high-throughput assays such as Next Generation Sequencing (*NGS*), gene-expression and SNP microarrays, enable the production of massive quantities of data that have to be opportunely stored to perform further analysis, such as discriminative statistical analysis [1], or data mining by using association rules [2, 3, 4] to find multiple correlations among SNPs.

Thus, researchers and clinician-researchers need databases able to store, and retrieve SNPs data. In addition, the integration of SNP data sets with clinical data, can be used for assessing disease risks, predict susceptibility, early diagnosing, and planning treatment regimens. Data collection, arrangement and storage are essential steps in the execution of daily activities ranging from the individual research level as well as in research areas. Hence, it is mandatory to permanently store and make data quickly accessible to the users through the employment of apposite containers such as databases. A database is a collection of data, used to manage and store information for any domain of interest.

Currently, genomics studies results are stored in flat textual files (not structured), and memorized locally into the research storage-system. The use of flat text files put some limitations such as: it makes difficult to integrate data referring to the same genomic study, to share data among researchers and research laboratories, excluding or making challenging the production of bigger and more informative datasets obtained by sharing and merging multiple SNP data sets. Thus, the need to represent not structured experimental genomics data in structured format arises. Goal that can be accomplished by mapping the experimental genomics information contained into the flat files in a structured model by means of a relational databases management system (*RDBMS*).

Thus, in this scenario where the data to handle are of different types, alongside the classic relational database management systems (RDBMS) [5], have been developed several different kinds of database management systems (*DBMS*) known as *NoSQL* (No Structured Query Language) [6]. *RDBMS* are relational data management systems based primarily on tables and use the *SQL* (Structured Query Language) language for data manipulation. Conversely, NoSQL databases are not relational data management systems, they are not based on tables to represent the association among data, and do not use SQL for data manipulation. Examples of NoSQL databases are graph databases, employed to handle data coming from social networks, biological networks (proteins interaction networks and biological pathways), where data do not require a relational model to be handled.

Databases have become essential for genetics, proteomics and interactomics studies.

The available public online SNP databases are a source of information from which researchers can readily retrieve and use data to design in silico studies. In silico studies allow researchers to improve the way to perform wet-lab experiments. Because, in silico studies make it possible to obtain in advance the most likely genomic regions affected by the SNPs to investigate. For

example, the knowledge of the genomic region to investigate avoids to the researchers to proceed by attempts saving in this way time and money, since a wet lab experiment may cost thousands of dollars.

There are different databases that provide information on SNPs data such as dbSNP [7], HGBASE [8], GBD [9], OMIM [10], and HapMap [11]. All the listed SNP databases represent a comprehensive, general and reliable catalogue of human SNPs, from which browsing and searching the core human genome SNP data. The listed databases contain relevant information for coding regions, including transcription structure, genome maps, bibliography, protein data, sequences, and the links to the supporting data, to facilitate large-scale studies in association genetics, pharmaco-genomics etc. On the other hand, the listed databases lack in the capability to handle experimental SNPs data obtained from clinical studies and enriched with clinical data. In this paper we present *snpStorage* a database able to collect experimental SNPs data obtained by using DMET (Drug Metabolism Enzymes and Transporters) [12] platform and enriched with clinical information. With respect to the above listed databases, snpStorage is not a comprehensive catalog of SNPs, but it is an attempt to produce a SNPs repository where research laboratories can store data (anonymously), merge and handle them, with which to yield more accurate SNP data sets annotated with clinical data when available, to perform more in-depth analysis. The goal pursuit by snpStorage is to make the analysis of integrated multiple SNP datasets more accurate w.r.t small SNP datasets, allowing users to extend and collect all their data in an unique repository. Current DMET SNPs outputs are arranged in a tabular form with 1931 rows (probes) and a number of column related to the number of patients enrolled in the study (commonly, in a number less than 100). Thus, snpStorage allows researchers of a research center or from different research centers, to store and query SNP data sets as well as, to merge together outcomes from several studies, making it possible to extend a small experimental study in a more informative one. Enlarging the size of an experimental study, allows to produce more informative data sets, providing a more accurate description of the study under-investigation, than analyzing data sets individually.

The rest of the paper is arranged as follows. Section 2 describes the major popular SNP databases, Section 3 describes the requirements analysis of snpStorage along with the input data. Section 4 illustrates how snpStorage has been implemented, highlighting the efficiency in store and retrieve information. Finally, Section 5 outlines the future directions of snpStorage and concludes the paper.

2. Related Work

Over the years, many public and private data repositories have been developed to facilitate the researcher's studies by freely collecting and sharing curated and inferred SNP data worldwide. Below, we report the list of some popular SNP databases.

- dbSNP [1] [7] is a comprehensive and reliable catalogue of human SNP variability with an efficient system to cross-reference multiple submissions of the same SNPs from centers outside NCB.

[1] https://www.ncbi.nlm.nih.gov/projects/SNP/

- The Human Gene Mutation Database (HGMD)[2] [13] is a catalog of disease-associated mutation. In particular, HGMD comprises the following information: SNPs in coding regions, regulatory and splicing of relevant regions in human, insertions/deletions, invariant regions and complex rearrangements.
- The Human Genic Bi-Allelic SEquences (HGBASE)[3] database [8] collects information about SNPs along with the connection among SNP and the affected gene. Data in HGBASE are obtained by collecting data from all the available public sources.
- The Genome Database (GDB)[4] [9] is a public data repository and catalogue of human's SNPs, genes and clones. GDB's data are obtained from the scientific literature and collecting submissions from genotyping centers.
- GenBank[5] [14] is a nucleotide sequence database available in NCBI. GenBank is a collection of sequences obtained from several different species updated daily. GenBank can provide to the users interested in SNP analysis specific context sequence of nucleotides to design a genotyping study.
- HapMap[6] [15] provides a wide range of data, but it is focused on SNPs data. HapMap allows users to graphically browse the genome to visualize the SNP variability currently available. Finally, SNPs characterized by HapMap are linked to their public databases from which retrieve more insight related to the specific entry.

All the listed databases pursue the goal of providing an online catalog containing all the identified genetic variations, which can be used to guide research in genomics, pharmacogenomics studies with information about the association of genetic variations and phenotypic traits. Conversely, from the listed database, snpStorage is not only a human SNP catalog. It provides users with a sandbox where they can store his/her data in an anonymous and secure place, encouraging collaboration among researchers who belong to the same research centers and affiliates to different research centers.

snpStorage's primary mission is to provide a collection of intuitive procedures that can automatize complex tasks like data harmonization, integration, and annotation, responsible for the difficulties encountered by the researcher in sharing their data to start new collaborations. In addition, to make it even more straightforward, using snpStorage for users unfamiliar with query languages, e.g., SQL, for data querying, and performing complex tasks such as harmonization, integration, and annotation, snpStorage will provide all these features interactively through a customized Graphical User Interface (GUI). In this manner, snpStorage could serve as a valuable resource for creating more accurate and helpful models to investigate complex conditions from a broader perspective, contributing to gain more knowledge than analyzing a single data set at the time.

Briefly, snpStorage is more than a repository where researchers can only store their experimental SNPs data sets obtained using genomics platforms (i.e., DMET microarrays). Still, it provides an automatic panel of graphical operations to quickly and intuitively enrich SNP with

[2] http://www.hgmd.cf.ac.uk/ac/index.php
[3] http://hgbase.cgr.ki.se
[4] http://www.gdb.org
[5] https://www.ncbi.nlm.nih.gov/genbank/
[6] https://www.genome.gov/10001688/international-hapmap-project

clinical information, joining several experimental SNP data sets, store the newly built data models, query new data set as well as old data sets, encouraging data sharing.

3. Requirement Analysis

Experimental SNPs data obtained from pharmacogenomics studies using DMET microarrays are arranged in a tabular format and saved as flat textual files (without a structure).

The current analysis practice poses some limitations, w.r.t. the possibility to increase the size of an experiment by merging more files regarding equivalent pharmacogenomics studies. Moreover, this merging operation has to be done manually by the researchers limiting at the same time the possibility to share data to obtain more informative data sets. Indeed, statistical analysis and machine learning techniques produce more accurate and relevant results when dealing with a huge amount of data than smaller ones. The current DMET outcomes are arranged in the form of a textual table saved in flat text files. In a generic DMET SNP data set, each row represents a probe (a well-defined genomic region known as chromosome containing several genes). In the current version of DMET microarray, there are 1931 probes able to investigate 233 genes involved in pharmacogenomics ADME (Absorption, Distribution, Metabolism, and Excretion). Instead, the number of columns is variable. The columns represent the number of patients enrolled in the experiments, the (*i-th*, *j-th*) cell contains the SNP detected in the *i-th* probes belonging to the *j-th* subject. Also, the subjects are memorized into the table using generic labels, making it impossible to harm the subjects' privacy. An example of DMET SNP dataset is depicted in Figure 1.

The possible operations that can be carried out on flat textual files are limited and complex because, researchers have to know how the data are related among them into the file to manually extract the information of interest to perform the next analysis. To figure out how the SNPs are distributed in a specific gene, for example the gene `CYP7B1` that consists of the following probes: `'AM_15053'`, `'AM_15085'`, `'AM_15086'`, and `'AM_15090'`, researchers have to extract for each subject the values related to those probes, to figure out the presence/absence of a specific SNP could be used as biomarkers for the condition under investigation. Gene extraction is not an easy operation due to the flat data representation of the DMET SNP data sets, so it must be performed manually. In addition, clinical data are stored in different formats and into separate flat text files, for which it is a not trivial operation to merge properly clinical and SNP data sets. The current structure and organization of SNP and clinical data sets highlights many limitations: *i)* do not allow to represent relationships among data; *ii)* make it challenging to share data among researchers into the same research center, as well among different research center; *iii)* limit the investigation power of the analysis. Mapping SNP data sets in a relational database system allows to overcome the actual limitation of flat files representation in particular: *i)* it makes it convenient to store and merge more experimental SNP and clinical data sets; *ii)* it promotes file sharing and consequently data integration; *iii)* it allows to perform further analysis other than classic case/control studies, such as it will enable to discriminate SNPs and clinical data in all the population. *iv)* Finally, it allows the extraction of complex relations among SNPs, diseases, and clinical condition quickly through queries.

4. snpStorage

To promote storing, merging, and sharing of experimental SNPs data sets with the possibility to integrate SNP data using clinical information, we implemented and developed snpStorage, a database using the relational database management systems (RDBMS) [5] MySQL server 8.0.13. snpStorage allows handling SNP data sets obtained by performing experimental investigation using the DMET microarray platform, and to integrate SNP data with clinical information if available. All the information related to subjects enrolled in a pharmacogenomics case-control

A	B	C	D	E	F	G	H
Probe Set ID	CEU_1	CEU_2	CEU_3	CEU_4	CEU_5	CEU_6	CEU_7
AM_10001	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10002	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10003	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10004	A/G	A/G	A/G	A/G	A/G	A/G	A/G
AM_10005	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10006	T/T	T/T	T/T	T/T	T/T	T/T	T/T
AM_10008	A/A	A/A	A/A	A/A	A/A	A/A	A/A
AM_10010	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10011	A/A	A/A	A/A	A/A	A/A	A/A	A/A
AM_10012	AG	G/G	G/G	G/G	G/G	G/G	G/G
AM_10013	A/A	A/A	A/A	A/A	A/A	A/A	A/A
AM_10014	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10016	T/T	T/T	T/T	T/T	T/T	T/T	T/T
AM_10017	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10019	A/A	A/A	A/A	A/A	A/A	A/A	A/A
AM_10020	T/T	C/T	C/T	C/T	C/T	C/T	C/T
AM_10021	A/A	A/A	A/A	A/A	A/A	A/A	A/A
AM_10022	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10023	T/T	T/T	T/T	T/T	T/T	T/T	T/T
AM_10024	A/A	A/A	A/A	A/A	A/A	A/A	A/A
AM_10025	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10028	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10030	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10031	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10033	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10034	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10035	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10039	A/T	A/T	T/T	T/T	T/T	T/T	T/T

Figure 1: Illustration of the outcomes obtained form a DMET microarray analysis.

study and stored in snpStorage are anonymized, making it impossible to harm patients' privacy. This is possible because snpStorage does not store sensitive information such as name, surname, address, etc., that could provide some clues on the identity of the subjects.

Analyzing Figure 1 it is worthy to note that the columns are named using generic labels, i.e., "CEU_1, CEU_2 ..." making it challenging to associate them with a real subject (except the data owner).

To store the SNP data sets permanently and to make it possible and convenient to extend the stored SNP data sets, it is necessary to map the flat textual SNP data sets into a database management system, along with all the additional, i.e., clinical information. In this way, we make it convenient and quick investigate, integrate, and share multiple SNP data sets. The tables used in snpStorage to map SNP data sets are: *Subjects* table contains the essential information necessary to link a subject with the healthcare center of belonging, the pathology under investigation, the detected SNPs, and the collected clinical data. *HealthcareCenters* table stores information about the center, the performed experiments, and the enrolled subjects in every study. *ClinicalInformation* table contains the clinical data, the pathology under investigation, and the information to link clinical data with the subject to which they belong. The pathology is stored and identified using the International Classification of Disease[7] (ICD). To overcome

[7]https://www.who.int/standards/classifications/classification-of-diseases

the main limitation imposed from the textual files, it is not enough to map the SNP data sets into the DBMS. It is necessary to relate the relationships between SNPs, diseases, subjects and laboratories to support complex queries. A possible solution is to group the 1931 probes using the genes and to link each gene to the *Subjects* table (let see Figure 2 gene table "*PPP*"). To accomplish the mapping task we used the annotations information provided by DMET microarray vendor Affymetrix [8]. As a result, we obtained 233 genes table (i.e., one table for each supported ADME gene from DMET platform), linking all the gene tables to the *Subject* table. Finally, the *ProbesGenesMapping* table contains the mapping between the probes and the gene of belonging, in the current version it contains only the DMET microarray mapping. Figure 2 shows the Entity-Relation schema (E/R) implemented to store and collect experimental DMET data enriched with clinical information. It is worthy to note that it is a simplified schema because we visualized 1 gene table (e.g., the PPP gene table) to simplify the visualization and readability of the diagram.

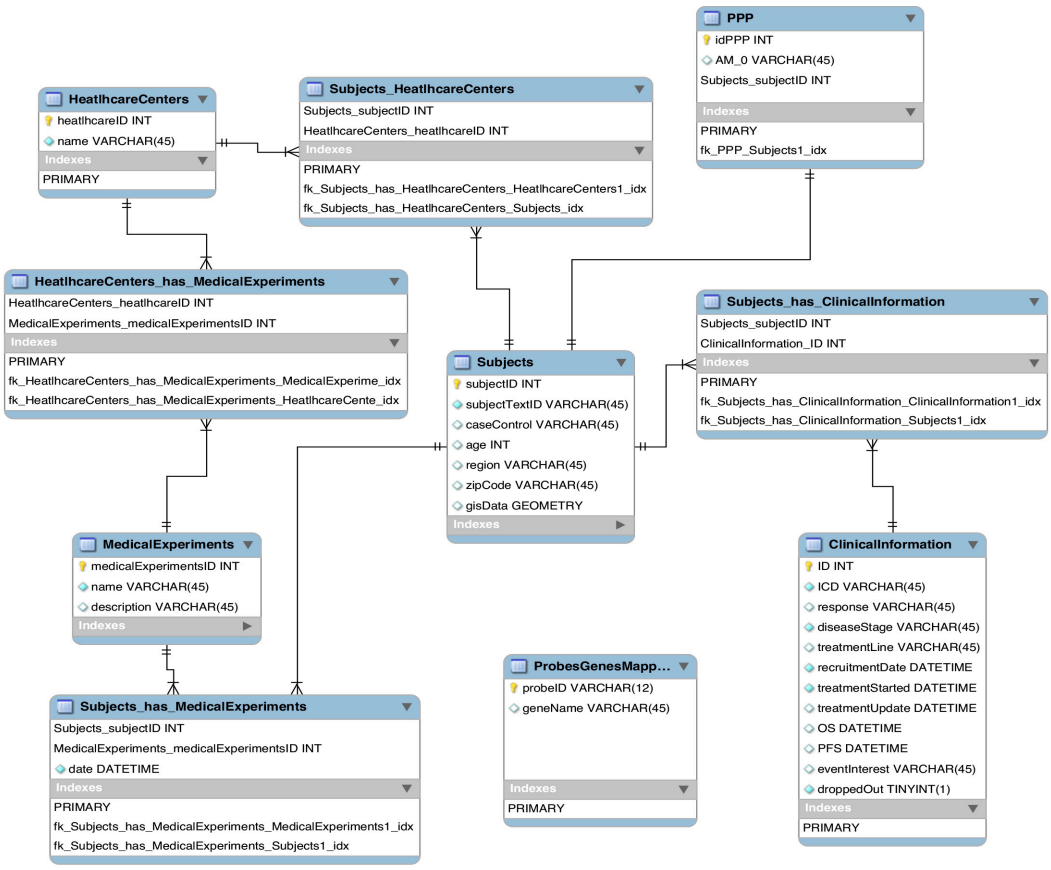

Figure 2: The E/R schema used to implement the data representation model.

snpStorage stores and efficiently handles SNPs and clinical information. Thus, it can help

[8] http://www.Affymetrix.com/support/technical/byproduct.affx?product=dmet_2

researchers to quickly share and integrate data from experimental pharmacogenomics SNP data sets possibly integrated with clinical data. Also, the integration of multiple SNP data sets, contribute to increasetypes of analysis to be performed, as well as the accuracy of the results. More data allow to describe the problem under investigation more accurately, opening new perspective not reachable by analyzing single small SNP data sets individually.

4.1. Performance evaluation

snpStorage is a relational database composed of 243 small tables. This design choice is due to the necessity to relate all the SNPs data with the correct probes, the diseases under investigation, the subjects and the laboratory/ies to which they belong. The relation schema allows obtaining narrow tables, with the advantage that they fit into the main memory. Thus the querying operation is faster than retrieving data from the secondary memory. Moreover, the SQL's operations such as *count*, *avg*, *sum* as well as *group by*, *order*, and so on, are available and can be computed easily through join operations among tables. The only limitation of this solution could be that join operations are CPU and RAM-intensive tasks. However, it is worthy to note that selecting some columns from specific tables does not require all the 243 sub-tables to be joined. Indeed, we will join only the sub-tables columns referenced in a given query.

An alternative solution to the multiple small tables could be the use of object-values columns. The object column type memorizes objects as values, allowing to obtain an enormous number of virtual columns into a single table. The object-value solution will enable tables to benefit from the compactness of regular columns while still allowing huge tables to exist, albeit less efficiently. Object columns require an extra layer of elaboration. The retrieved data have to be converted into the native data type, introducing extra CPU work to parse the object in the native type object. The object approach's advantage is that joining operations among tables is unnecessary since whole data can be held into a single table.

To compare the two possible implementations, we created small and object tables, evaluating the query's average response time by simulating multiple parallel queries using multiple parallel threads. Both tables contain the same number of tuples equals to $1,000$. The average response times refer to the time necessary to perform the queries on both implementation the small and object tables, respectively. As a benchmark, we used two queries, a simple *Select* query and a *Count* query, monitoring even if the number of simultaneously performed queries influences the average query time. The *select* queries are: `SELECT * FROM Subjects_smallTable` and `SELECT * FROM Subjects_objectColumn`. The *count* queries are: `SELECT COUNT(Allele) FROM Subjects_smallTable WHERE Allele = "A/A";` and `SELECT * FROM Subjects_objectColumn`. To count in the objectColumn table we need to select all the value and after convert the object in the native domain and counting it by using the additional computational layer. The query times on both representation are reported in Table 1 and Table 2. Analyzing Figure 3 we can see that until the queries retrieve values without being necessary to manipulate it, the trend querying between small and object tables is quite similar. Running a simple *count* query on both models, we can see a slight increase in the response query time in small table querying (since it added a little extra work-load due to the count function). Conversely, the response query's time increase in some order of magnitude when necessary to use the additional layer of computation to extract the further essential information that returns the count value of the

submitted query.

In conclusion, snpStorage allows to convert unstrutured SNP data sets in structured data sets, as well as to take advantage of the native SQL functions to handle the stored data. In this way, further complex data analysis can be performed faster and in a shorter time since it can handle directly relevant data without using an additional computational layer to extract the necessary information.

Table 1
The measured average time in second necessary to perform the *Select* and *Count* query on the database containing small tables implementation, varying the number of threads used.

	#Th=1	#Th=5	#Th=10	#Th=15	#Th=20
Select	4.218	6.36	6.571	8.24	8.105
Count	4.383	5.957	6.29	8.009	9.751

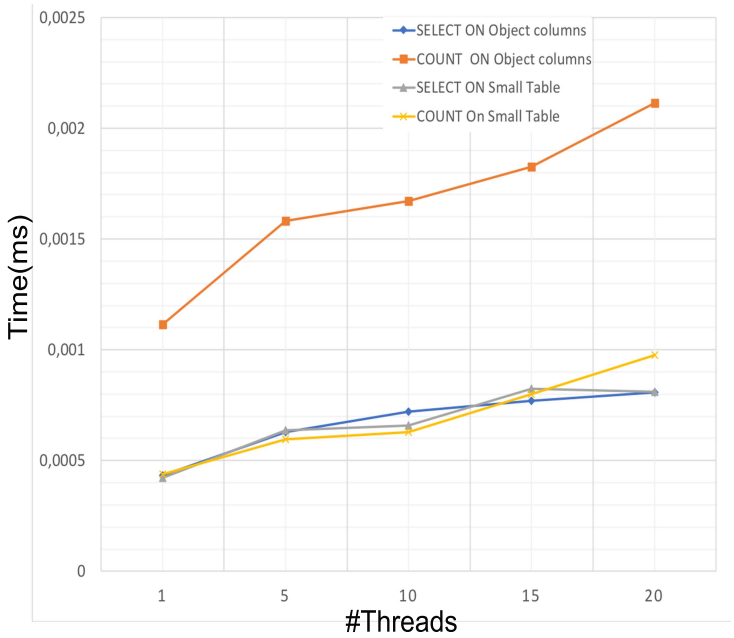

Figure 3: Average times of multiple *Select* and *Count* queries.

Table 2
The measured average time in second necessary to perform the *Select* and *Count* query on the database based on object column tables implementation, varying the number of threads used.

	#Th=1	#Th=5	#Th=10	#Th=15	#Th=20
Select	4.319	6.277	7.203	8.070	7.477
Count	11.147	15.809	16.721	18.259	21.1344

5. Conclusion

SNPs are a relevant source of information spanning from genomics to pharmacogenomics and personalized medicine. Linking SNP information with clinical data allows spurring light on the possible hidden relation between the healthy/diseased status of subjects and the arrangement of the patient's genome. In this way, SNPs may help scientists assume better the susceptibility of an individual to a disease or the adverse drug reaction. Since snpStorage stores and efficiently handles SNPs and clinical information, it can help researchers to quickly discriminate information among SNPs and clinical data into the population under investigation through simple queries. Future works will be focused on developing and implementing the web interface of snpStorage, through which analysis pipelines will be made available to help researchers quickly analyze their data stored into snpStorage without being necessary to use additional software tools. snpStorage pursues the target to provide an environment where researchers can perform queries built in a graphical way, without to be necessary to use SQL, on specific genes a group of genes, or consider detailed clinical information. Thus, the unification of the SNPs and clinical data represents a step toward defining individual and personalized treatments of complex diseases such as psychiatric disorders, diabetes, hypertension, and cancer.

Finally, snpStorage is more than a repository where researchers can only store their experimental SNPs data sets obtained using genomics platforms (i.e., DMET microarrays). Still, it provides an automatic panel of graphical operations to quickly and intuitively enrich SNP with clinical information, joining several experimental SNP data sets, store the newly built data models, query new data set as well as old data sets, encouraging data sharing.

References

[1] P. H. Guzzi, G. Agapito, M. T. Di Martino, M. Arbitrio, P. Tassone, P. Tagliaferri, M. Cannataro, Dmet-analyzer: automatic analysis of affymetrix dmet data, BMC Bioinformatics 13 (2012) 258. URL: https://doi.org/10.1186/1471-2105-13-258. doi:10.1186/1471-2105-13-258.

[2] G. Agapito, P. H. Guzzi, M. Cannataro, Dmet-miner: Efficient discovery of association rules from pharmacogenomic data, Journal of Biomedical Informatics 56 (2015) 273 – 283. URL: http://www.sciencedirect.com/science/article/pii/S153204641500115X. doi:https://doi.org/10.1016/j.jbi.2015.06.005.

[3] G. Agapito, P. H. Guzzi, M. Cannataro, Parallel extraction of association rules from genomics data, Applied Mathematics and Computation 350 (2019) 434–446.

[4] G. Agapito, P. H. Guzzi, M. Cannataro, Parallel and distributed association rule mining in life science: A novel parallel algorithm to mine genomics data, Information Sciences (2018).

[5] P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, T. G. Price, Access path selection in a relational database management system, in: Proceedings of the 1979 ACM SIGMOD International Conference on Management of Data, SIGMOD '79, ACM, New York, NY, USA, 1979, pp. 23–34. URL: http://doi.acm.org/10.1145/582095.582099. doi:10.1145/582095.582099.

[6] J. Han, H. E, G. Le, J. Du, Survey on nosql database, in: 2011 6th International Conference

on Pervasive Computing and Applications, 2011, pp. 363–366. doi:10.1109/ICPCA.2011.6106531.

[7] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, K. Sirotkin, dbsnp: the ncbi database of genetic variation, Nucleic Acids Research 29 (2001) 308–311. URL: http://dx.doi.org/10.1093/nar/29.1.308. doi:10.1093/nar/29.1.308.

[8] A. J. Brookes, H. Lehväslaiho, M. Siegfried, J. G. Boehm, Y. P. Yuan, C. M. Sarkar, P. Bork, F. Ortigao, Hgbase: a database of snps and other variations in and around human genes, Nucleic Acids Research 28 (2000) 356–360. URL: http://dx.doi.org/10.1093/nar/28.1.356. doi:10.1093/nar/28.1.356.

[9] S. I. Letovsky, R. W. Cottingham, C. J. Porter, P. W. D. Li, Gdb: The human genome database, Nucleic Acids Research 26 (1998) 94–99. URL: http://dx.doi.org/10.1093/nar/26.1.94. doi:10.1093/nar/26.1.94.

[10] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders, Nucleic Acids Research 33 (2005) D514–D517. URL: http://dx.doi.org/10.1093/nar/gki033. doi:10.1093/nar/gki033.

[11] G. A. Thorisson, A. V. Smith, L. Krishnan, L. D. Stein, The international hapmap project web site, Genome Research 15 (2005) 1592–1593. URL: http://genome.cshlp.org/content/15/11/1592.abstract. doi:10.1101/gr.4413105. arXiv:http://genome.cshlp.org/content/15/11/1592.full.pdf+html.

[12] M. Arbitrio, M. T. Di Martino, F. Scionti, G. Agapito, P. H. Guzzi, M. Cannataro, P. Tassone, P. Tagliaferri, Dmet™(drug metabolism enzymes and transporters): a pharmacogenomic platform for precision medicine, Oncotarget 7 (2016) 54028–54050. URL: https://www.ncbi.nlm.nih.gov/pubmed/27304055. doi:10.18632/oncotarget.9927.

[13] P. D. Stenson, M. Mort, E. V. Ball, K. Howells, A. D. Phillips, N. S. Thomas, D. N. Cooper, The human gene mutation database: 2008 update, Genome Medicine 1 (2009) 13. URL: https://doi.org/10.1186/gm13. doi:10.1186/gm13.

[14] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers, GenBank, Nucleic Acids Research 41 (2012) D36–D42. URL: https://doi.org/10.1093/nar/gks1195. doi:10.1093/nar/gks1195. arXiv:https://academic.oup.com/nar/article-pdf/41/D1/D36/3680750/gks1195.pdf.

[15] I. H. Consortium, et al., A second generation human haplotype map of over 3.1 million snps, Nature 449 (2007) 851.