

Article

A parallel software pipeline for personalized medicine

Giuseppe Agapito ¹ , Pietro Hiram Guzzi ¹  and Mario Cannataro ^{1,*} 

¹ Data Analytics Research Center, Department of Medical and Surgical Sciences, University "Magna Græcia" of Catanzaro, Viale Europa, 88100 Catanzaro, Italy; {agapito, hguzzi, cannataro}@unicz.it

* Correspondence: cannataro@unicz.it; Tel.: +39-0961-3694100

Abstract: Personalized medicine is an aspect of the P4 medicine (predictive, preventive, personalized and participatory) based precisely on the customization of all medical characters of each subject. In personalized medicine, the development of medical treatments and drugs is tailored to the individual characteristics and needs of each subject, according to the study of diseases at different scales from genotype to phenotype scale. To make concrete the goal of personalized medicine, it is necessary to employ high-throughput methodologies such as Next Generation Sequencing (NGS), Genome-Wide Association Studies (GWAS), Mass Spectrometry or Microarrays, that are able to investigate a single disease from a broader perspective. For example, by using genotyping microarrays (e.g. collections of Single Nucleotide Polymorphism - SNP) it is possible to uncover the reasons (i.e. mutation in genes) because a treatment works properly in some patients (for example absence of mutated genes), but it does not work (presence of mutated genes) in others. A side effect of high-throughput methodologies is the massive amount of data produced for each single experiment, that poses several challenges (e.g. high execution time and required memory) to bioinformatic software. Thus a main requirement of modern bioinformatic software is the use of good software engineering methods and efficient programming techniques, able to face those challenges, that include the use of parallel programming and efficient and compact data structures. Thus, to exploit all the potential of this massive amount of data in the short possible time (before that data becomes obsolete), the necessity to develop parallel software tools for efficient data collection and analysis arise. Moreover, due to the heterogeneity of the data produced by the different kinds of experimental platforms, it is necessary to automatize in a comprehensive software pipeline, the various steps that compose a bioinformatic analysis, such as: the preprocessing of raw data to remove noise or corrupted data; the annotation of data with external knowledge (e.g. Gene Ontology), and the integration of molecular data with clinical data. It should be noted that such steps are necessary to make statistical or data mining analysis more effective. This paper presents the design and the experimentation of a comprehensive software pipeline, named microPipe, for the preprocessing, annotation and analysis of microarray-based SNP genotyping data. A case study in pharmacogenomics is presented. The main advantages of using microPipe are: the reduction of errors that may happen when trying to make data compatible among different tools; the possibility to analyze in parallel huge datasets; the easy annotation and integration of data.

Keywords: SNP; Multiple Analysis Pipeline; Pharmacogenomics; Overall Survival Curves; Data mining; Statistical Analysis.

1. Introduction

The continuous improvements in experimental technologies allow spreading the use of genotyping analysis in several biological, medicals and clinical areas [1,2]. The investigation of complex diseases such as cancer, alzheimer, and leukemia, requires to investigate multiple actors (i.e., genes, proteins and small molecules) at the same time. Thus, it is necessary to use high-throughput

experimental platforms such as Next Generation Sequencing (NGS), Genome-Wide Association Studies (GWAS), Mass Spectrometry and Microarrays. GWAS offers a large-scale genotyping of SNPs in thousands of DNA samples and it is one of the most powerful methods for recognizing genes related to complex diseases as reported in [3]. SNP microarrays have been used to identify the genetic polymorphisms involved in drug transporters and metabolizing enzymes that play, for instance a crucial role in the interindividual variability concerning efficacy and toxicity of erlotinib treatment [4]. In this studies [5,6], the effectiveness of using DMET microarray in a clinical trial have been proved.

A side effect of high-throughput methodologies is the massive amount of data produced for every single experiment, which poses several challenges regarding execution time and required memory. The format used to code the experimental data is not suitable to be directly used as input with the most used data mining and statistical software available, e.g., Weka, SPSS, R and many others. Thus a lot of efforts to make datasets compatible with the chosen tool are demanded to the researchers in order to make it possible to extract actionable knowledge from these amount of data. Indeed, high-throughput vendors provide only basic software for the analysis of raw experimental data, in order to produce data that can be used for further data mining and/or statistical analysis.

To carry out data mining and/or statistical analysis on an input dataset, researcher has to manually perform data manipulation to make the dataset in a suitable format to be used as input for the chosen tool. For example, researchers to conduct a statistical analysis with SPSS or R by using as input a DMET SNP dataset, have to convert the literals SNPs contained into the dataset in numbers to be compatible with R and SPSS to start the analysis. This conversion is mandatory and has to be done manually. No tools are available to guide the users with basic or none programming skills, making conversion a tedious and error-prone task. A fast way for users with advanced programming skills is to write a script for each tool and use it to convert the input datasets. This is only half of the battle, indeed, to analyze the same dataset from a data mining perspective, it is not possible to use the dataset converted for the statistical analysis, but it is necessary to manipulate the original dataset. For example, extracting association rules from a DMET SNP dataset by using Weka, researchers have to transpose and convert in the ARFF format the dataset. DMET microarrays are employed to investigate polymorphic variants related to Absorption, Distribution, Metabolism, and Excretion (ADME) genes significantly contributing to individual patients' drug sensitivity, resistance, and toxicity [7]. DMET SNP datasets are arranged like a huge table (rows identify probes, columns identify subjects under investigation), the transposition operation consists in translating rows in columns, producing a more suitable data format from which extract association rule (i.e., the current real DMET datasets have 1931 rows, but a variable number of subjects). Also, users by using Weka are forced to use Apriori algorithm, because the PF-Growth algorithm in Weka works only with dichotomous transactions, this is each row has to contain not more than two different elements.

Thus, the primary requirement of modern bioinformatic software is the use of suitable software engineering methods and efficient programming techniques, able to face those challenges, which include the use of parallel programming and efficient and compact data structures as well as the automatic management of tedious steps such as preprocessing, summarization and so on. Besides, it is still essential to provide domain tools, especially designed for those scientists with basic or none programming skills, or even with limited time, that look for straightforward solutions with low time investment.

To alleviate the researchers work in a past work we developed DMET-Analyzer [8] to automatically analyze a whole DMET SNP datasets in order to figure out the most relevant single SNP, i.e., related with the adverse drug reaction. On the other hand, to extract multiple SNPs contemporarily related to the adverse drug reactions from DMET SNP datasets, we have developed DMET-Miner [9]. Moreover, to make DMET SNP dataset useful in clinical practice, DMET datasets can be annotated with clinical data such as overall survival (OS), progression-free survival (PFS), metastasis, and so on. Thus, to analyze such annotated data, we developed OS-Analyzer [10]. OS-Analyzer can analyze a whole DMET dataset annotated with the temporal events, and results in probes sorted by statistical relevance.

In this way, OS-Analyzer allows the researchers to graphically discriminate which SNP is related to a good or bad overall survival.

To further speed up and simplify the analysis of DMET SNP microarrays we present *microPipe* a novel tool to perform on the same input dataset statistical, survival and data mining analysis. *microPipe* is a tool able to perform OS and PFS analysis, data mining of association rules and Fisher discriminant analysis, starting from a DMET SNP dataset annotated with temporal events. The only operation required for the users is to load the input dataset. In summary, *microPipe* is a software environment that integrates the functions of DMET-Analyzer, DMET-Miner and OS-Analyzer.

The remainder of the manuscript is structured as follows: Section 2 describes previous and related work to the DMET analysis tools. Section 3 presents *microPipe*, Section 4 presents a case study employing *microPipe*. Finally, Section 5 concludes the manuscript.

2. Related Work

Recently, microarrays have become one of the most suitable methodologies in genomic research, due to the high magnitude of simultaneously analyzed genes per single experiment. Microarrays are involved in many life science areas, spurring for the development of suitable data analysis models and tools. On the other hand, the necessity to provide researchers software tools simple to use, arises as well.

In this Section, we summarize the software tools listed on the OMICtools¹ website and evaluated as compatible with DMET data, and the software tools able to deal with DMET SNP microarray datasets developed in our research group. We assessed tools that can extract knowledge from DMET SNP datasets and tools that can do only modelling and prediction of the structure of DMET enzyme.

- DMET-Analyzer [8] is a tool for the automatic association analysis of the difference of the patient genomes and the clinical status of patients, e.g., the different response to drugs. The current version of DMET-Analyzer verifies, for each SNP, the association between the presence of SNP and the classes of patients (e.g. RESP or NOTRESP) yet determined through the use of the well known Fisher's test. Two multiple test corrections are available (Bonferroni and False Discovery Rate) to improve the statistical significance of results. Finally, DMET-Analyzer annotates significant SNPs with information provided by Affymetrix libraries and with links to the dbSNP database (for basic information about SNPs) and the PharmGKB pharmacogenomics knowledge base, giving various information (e.g., pathways) related to pharmacogenomics.
- DMET-Miner [9] is a software tool for the automatic extraction of association rules that correlate the presence of a series of allelic variants with the clinical status of patients, for example, the combination of alleles typical of a class responsible for the different response to drugs. DMET-Miner allows users to calculate automatically and easily association rules from DMET datasets.
- OS-Analyzer (OSA) [10] is a software tool for the computation and visualization of Overall Survival (OS) and Progression Free Survival (PFS) curves of cancer patients and evaluate their association with ADME gene variants. OS-Analyzer can perform an automatic analysis of DMET data enriched with survival events. The ranking of the results is made according to statistical significance obtained by comparing the area under the curves by using the log-rank test, allowing a quick and easy analysis and visualization of high-throughput data.
- coreSNP [11] is a tool for parallel pre-processing and statistical analysis of DMET datasets. The scalable implementation based on multi-threading allows coreSNP to manage vast volumes of pharmacogenomic experimental data. The automatic association analysis of the variation of the patient genomes and the clinical conditions of patients (e.g., the different response to drugs), is computed by implementing the well known Fisher's test. Moreover, multiple-test

¹ <https://omictools.com>

corrections (Bonferroni and False Discovery Rate) to improve the statistical significance of results, are available. The visualization of the SNPs detected into the entire dataset as a heat map, give immediate visual feedback, allowing users to interpret the results quickly. coreSNP is the parallel version of DMET-Analyzer.

- PARES (Parallel Association Rules Extractor from SNPs) [12] is a multi-thread software tool for the parallel extraction of association rules. PARES is a multi-thread version of an optimized version of the Frequent Pattern Growth (FP-Growth) algorithm. PARES is the multithread version of DMET-Miner. PARES thanks to the simple and intuitive graphic user interface, it is a software tool easy to use, where, specific skills are not necessary to easily extract multiple relations between genomic factors buried into the data sets.
- *Cloud4SNP* [13] is a Cloud-based bioinformatics tool for the parallel preprocessing and statistical analysis of pharmacogenomics SNP DMET microarray data. It is a Cloud-based version of *DMET-Analyzer* that has been implemented on the Cloud using the Data Mining Cloud Framework [14], a software environment for the design and execution of knowledge discovery workflows on the Cloud [15]. It allows to statistically test the significance of the presence of SNPs in two classes of samples using the well known Fisher test. *Cloud4SNP* uses data parallelism and employs an optimization technique to avoid the execution of useless Fisher tests, through the filtering of probes with similar SNPs distributions.
- ADMET Predictor is a software tool to predict the 2D structure of the molecule to create high-quality models with which produce compounds. Molecules' structure is done through a graphical user interface with which manipulate and visualize data. ADMET Predictor is available for Unix/Linux, Mac OS, Windows operating systems. ADMET Predictor requires the purchase of a license to be used. An evaluation version of the software can be obtained by full fill a request form available on the ADMET Predictor's website <http://www.simulations-plus.com/software/admetpredictor/>.
- ADMET Descriptors allows assessing compounds through queries. ADMET Descriptors is part of the BIOVIA Discovery Studio software. ADMET Descriptors provides to the user through GUI function to calculate the predicted absorption, distribution, metabolism, excretion and toxicity (ADMET) properties for a collection of molecules. ADMET Descriptors is available for Unix/Linux, Mac OS, Windows operating systems. To download a trial version, users have to fill the online form. The download of the complete version of the software requires purchasing a license, and can be done at the following web address <http://accelrys.com/products/collaborative-science/biovia-discovery-studio/qsar-admet-and-predictive-toxicology.html>.
- PreADMET is a web application for predicting ADME data and producing drug-like library using in-silico methods. PreADMET is a web application and can be freely accessed by a browser at the following web address <https://preadmet.bmdrc.kr>. PreADMET is also available in purchasing editions. PreADMET consists of four components Molecular Descriptor Calculation, Drug-likeness Prediction, ADME Prediction, and Toxicity prediction. PreADMET is developed by using C and PHP programming languages.
- ADMEWORKS ModelBuilder is a tool dedicated to build models that can later be used to predict different chemical and biological properties of compounds. ADMEWORKS ModelBuilder provides quantitative/qualitative structure-activity relation's analysis, multiple statistical methods for generating predictive models, interactive graphical feature and outlier selection tools and automated analytical tools for feature and sample selection. ADMEWORKS ModelBuilder requires to purchase a license; trial version download is possible only filling the registration form. ADMEWORKS ModelBuilder is available at the following web address <http://www.fqs.pl/en/chemistry/products/adneworks-modelbuilder>.
- ADMEWORKS Predictor is a Web-based application to evaluate the ADMET properties of compounds, available at <http://www.fqs.pl/en/chemistry/products/adneworks-predictor> web site. ADMEWORKS Predictor provides interactive graphical features through a web browser,

view structures using a highly functional 3D structure viewer. Store molecular structures and properties in the database, and function for filtering and sorting data.

Table 1. Comparison of the functionality provided by the tools compatible with DMET reviewed on OMICtools website. In the table, OS refers to the Operating System compatibility. *WUML* is short for Windows, Unix, Mac, and Linux. *na* means that the information was not available. *WebApp* refers to a web application that can be used through a web browser. The ✓ symbol indicates that the tool provides that function, whereas × means an unsupported feature.

TOOL	OS	Statistical	Data Mining	Prediction	Modelling	Free	Pay
DMET-Analyzer	WUML	✓	×	×	×	✓	×
DMET-Miner	WUML	×	✓	×	×	✓	×
OS-Analyzer	WUML	✓	×	×	×	✓	×
coreSNP	WUML	✓	×	×	×	✓	×
PARES	WUML	×	✓	×	×	✓	×
cloud4SNP	WUML	✓	×	×	×	✓	×
ADMET Predictor	WUML	×	×	✓	×	×	✓
ADMET Descriptors	WUML	×	×	✓	×	×	✓
PreADMET	WebApp	×	×	✓	×	×	✓
ADMEWORKSModelB	na	×	×	✓	✓	×	✓
ADMEWORKSPredictor	WebApp	×	×	✓	×	×	✓

It is worthy to note that, only the first six tool in the table can extract actionable knowledge hidden into DMET SNP datasets. The other tools although they refer to DMET are not able to analyze DMET SNP dataset but can be employed only to conduct modeling and prediction analysis of the structure of molecules related to DMET.

3. Results

To help researchers to perform survival, data mining and statistical analysis in an easy and fast way, we developed and implemented microPipe. microPipe is a software tool implemented by using Java 8.0 language. The only requirement to use microPipe on users' machine is that Java 8 or higher is already installed. Otherwise, it is mandatory to install Java and after run microPipe. microPipe has been thought to be easy to use by users with basic or none programming skills through an interactive Graphical User Interface (GUI), as well as to be scalable even when analyzing huge data sets.

microPipe is a fast, efficient, and portable tool designed to assist the users in statistical analysis, in Kaplan-Meier curve computing/visualization, and in association rules mining from DMET SNP datasets. microPipe comes with an advanced computational engine, which allows to perform data analysis in parallel by taking advantage of the available hardware. microPipe's engine is designed to exploit the multi-CPU's multi-Cores architectures to speed-up the computational analysis. microPipe automatically maps each phase of the analysis process to a single Core/CPU, and if the available number of available Cores/CPU's. Thus, each analysis steps can be done in parallel by exploiting multithread computation, without saturating the system, through some programming methods.

The microPipe's architecture is designed to be efficient as well as simple to be extended by adding new components. The architecture is depicted in Figure 1. Every new component that has been developed implementing the microPipe *outlet-interface*, can be added by copying it in the microPipe *AppContainer*. The main modules of microPipe's architecture are: *i) FileLoader* loads and checks the consistency of input file. *ii) FileParser* parses the loaded file, making the data suitable for the next phases. *iii) FilePreprocess* cleans and deals with data to make data in a format suitable for the application. *iv) AppDispatcher* receives the preprocessed data and dynamically binds the data with the proper application. *v) ThreadsMapper* starts the analysis pipeline, where each application can be execute in parallel and independent of the others. *vi) GUIEventManager* is the graphical control that allow to microPipe and users to interact among them through a simple and intuitive graphical commands.

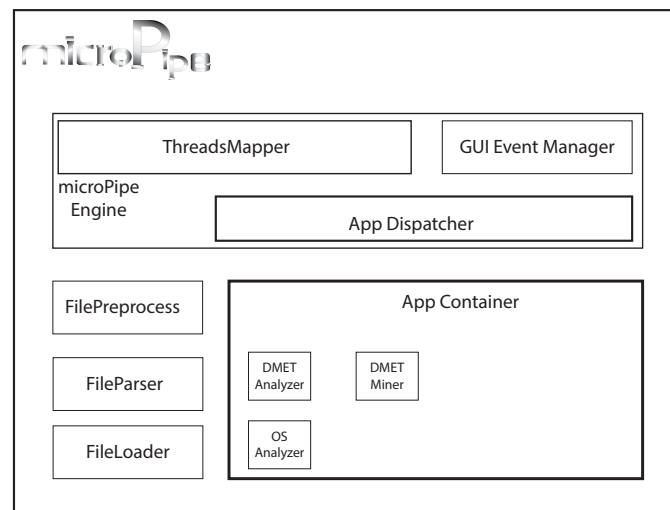


Figure 1. The microPipe's architecture.

The microPipe primary function is to avoid that users have to use multiple tools to detect and identify the relevant SNPs involved in the different response of subjects to the treatment. microPipe can produce heatmap SNPs visualization, designed to quickly highlight the SNPs' distribution among the subjects involved in the study. Using the well known Fisher's Test it can automatically highlight the variation of the patient genomes and the clinical conditions of patients, e.g., the different response to drugs. Moreover, users by selecting a statistical corrector between Bonferroni and False Discovery Rate (FDR) can produce more accurate results, reducing the bias due to the *type I errors*. microPipe association rule mining is done automatically and in parallel with the other analysis. microPipe demands to the user the setting of parameters for the FisherFilter's value with which microPipe prunes useless rows, reducing the search space. The setting up of confidence and minimum support values, allows microPipe to show to the user only the relevant association rules that satisfy both values and that are more relevant. Finally, microPipe shows to the users the probes sorted by statistical relevance obtained computing the log-rank values. In this mode user can analyze and save only significant survival curves.

The main feature of microPipe is the capability to perform on the same DMET SNP dataset annotated with temporal events, multiple type of analysis automatically. Moreover, microPipe automatically preprocesses the data through a suitable methodology able to remove useless data, i.e., rows with low information content, and avoiding that users have to repeat manually tedious configuration settings to figure out important clues.

4. Case Study

In this Section, we describe an use case involving microPipe. The proposed use case highlights remarkably, the main steps that a user has to follow to get actionable knowledge from a real DMET SNP dataset. As the first step, the user has to run microPipe. microPipe can be started by double click on the file "microPipe.jar," modality recommended for a user with basic programming skills, or by using the command line. From the command line, expert users can set up specific features of the Java virtual machine, e.g., the heap size, the stack size, to better customize the execution environment. "microPipe.jar" is distributed as Java archive (jar) tool provided as part of the Java Development Kit (JDK) a compressed and executable file. To execute a program distributed as JAR package from the command line, users have to type "\$ java -jar microPipe.jar". When started microPipe shows the main window (see Figure 2).

microPipe presents an essential and minimal graphical users interface (GUI). The GUI comprises only the menu "File" (see Figure 2) containing the Load and exit commands. By clicking on "exit"



Figure 2. The microPipe’s main window.

closes microPipe. Instead, by clicking on "Load" option, it will be visualized the file chooser window, with which user can browse his/her hard disk to load the input file to analyze. After located and selected the file, by clicking the "ok" button, microPipe automatically starts the analysis. Each step of the analysis pipeline s done in parallel and independently of each other, we the data mining and statistical analysis are done interactively to get some user feedback to produce relevant results. Conversely, survival analysis is done in background because it is not required input feedback from the user.

As first step of the analysis pipeline, microPipe will provide the "OS Navigation Panel Results" and "PFS Navigation Panel Results" windows see (Figure 3), that are the output of survival analysis phase.

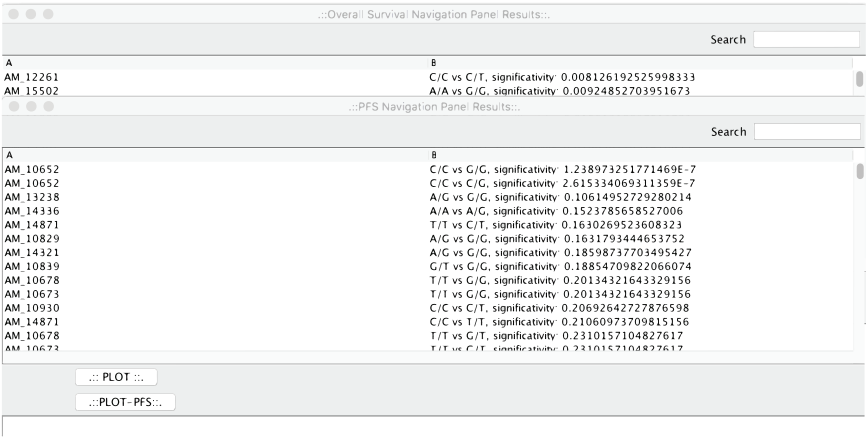


Figure 3. The OS and PFS navigation panels. From that panels, the user can select the most relevant curves to be plotted.

The two windows contain the results sorted by statistical relevance obtained by using the log-rank test. In this way, microPipe provides to the user a complete view of the most suitable probes that should be further investigated. Conversely from other tools like SPSS, or R, where users have to configure the analysis of the whole dataset manually, by using microPipe all the repetitive configuration steps are done by the software. Thus, users analyze the relevant curves by clicking on each curve and than visualize the related OS or PFS or both curves. If necessary, this chart can be saved on the disk as an image. Moreover, on the bottom of the chart, the medians values and Hazard Ratio Value of each curve are presented, see (Figure 4), highlighting the key features speeding up the understanding of the OS curve.

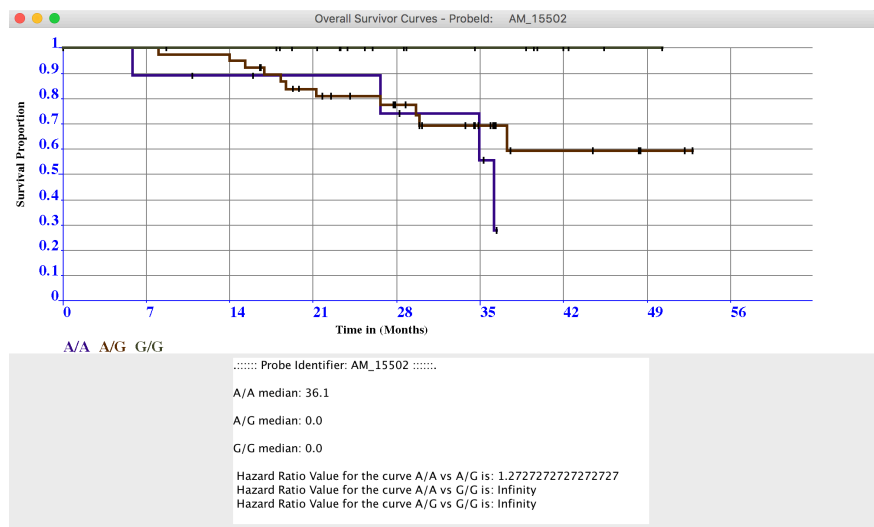


Figure 4. In Figure is presented the OS curve related to the probe AM_15502, obtained by plotting the data in the OS Navigation Panel by clicking on it.

Depending on the complexity of the input dataset microPipe warns users to select the subject belonging to class B (see Figure 5). In this way, microPipe signals to the users that the statistical and data mining phases are starting.

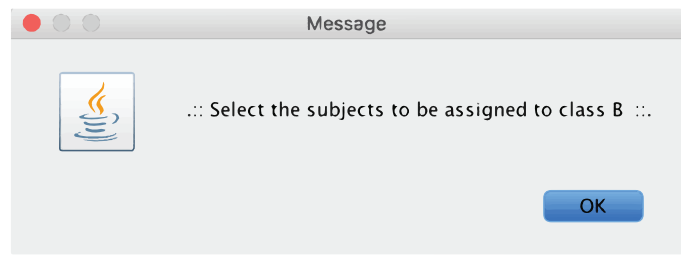


Figure 5. The subjects' selection message.

After assignment of subjects is done, by clicking on the *preprocess* button (see Figure 6), the user starts statistical and data mining analysis.

As intermediate step microPipe will show the heat-map of SNPs' distribution, and the preprocessed table obtained by filtering useless rows see (Figure 7).

From the menu bar in Figure 8, users can perform the Fisher's Test on all the pre-processed table, or chose to compute the Fisher's Test on a specific probe or to calculate the Hardy-Weinberg equilibrium.

Choosing to compute "exhaustiveFisherTest" with or without correctors, all the relevant Fisher test are conveyed see (Figure 9).

The results of Fisher tests allow users to discover SNPs related to the particular disease, or treatment under investigation (see Figure 9). As the final step of the analysis pipeline, microPipe will inform the user that it is setting up association rules mining see Figure (10).

microPipe will ask users to choose the FisherFilter value that is different from that used in the statistical analysis (see Figure 11) with which to remove as many as rows possible that are under the chosen threshold value.

The filtering value is automatically computed for every single row (probe). The setting up of confidence and minimum support values, allows microPipe to show to the user only the strong association rules that at least meet both values (see Figure 12).

DMET-DatasetPlusOS.txt :: DMET-Analyzer: Table of Probes & Samples ... /Users/giuseppagapito/D...

A	B	C	D	E	F	G	H	I	J
PROBEID	NONRESP	NONRESP	NONRESP	NONRESP	NONRESP	NONRESP	NONRESP	NONRESP	NONRESP
AM_10001	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10003	T/T	C/T	C/T	T/T	C/T	C/T	C/C	C/C	C/C
AM_10004	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10005	T/T	C/T	C/T	T/T	C/T	C/T	C/C	C/C	C/C
AM_10006	T/T	C/T	C/T	T/T	C/T	C/T	C/C	C/C	C/C
AM_10010	C/C	C/T	C/T	C/C	C/T	C/T	T/T	T/T	T/T
AM_10011	A/A	A/G	A/G	A/A	A/G	A/G	G/G	G/G	G/G
AM_10013	A/A	A/G	A/G	A/A	A/G	A/G	G/G	G/G	G/G
AM_10014	C/C	C/T	C/T	C/C	C/T	C/T	T/T	T/T	T/T
AM_10016	T/T	C/T	C/T	T/T	C/T	C/T	C/C	C/C	C/C
AM_10020	C/C	C/T	T/T	C/C	C/C	C/T	T/T	C/C	C/C
AM_10021	A/A	A/A	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	G/G	G/G
AM_10023	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/T
AM_10024	A/A	A/G	G/G	A/A	A/A	A/G	G/G	A/A	A/A
AM_10025	A/A	A/G	G/G	A/A	A/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	C/C	C/C
AM_10031	T/T	G/G	T/T	G/G	G/G	G/G	G/G	G/T	G/T
AM_10033	G/G	G/G	C/C	A/A	A/G	G/G	C/C	G/G	G/G
AM_10034	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10035	T/T	G/G	T/T	G/G	G/G	G/G	G/T	G/T	G/T
AM_10047	C/C	C/C	C/C	C/C	C/C	C/T	C/C	C/T	C/C
AM_10048	T/T	T/T	C/T	C/C	T/T	C/C	C/T	C/T	C/C
AM_10052	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10053	C/C	C/C	C/C	C/T	C/C	C/T	C/C	C/T	C/C
AM_10054	A/A	A/G	A/A	A/A	A/A	A/A	A/A	A/A	A/A
AM_10056	A/A	NOCALL	A/A	A/A	NOCALL	A/A	A/A	A/A	A/A
AM_10059	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10064	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10070	G/G	G/G	G/G	G/G	G/G	A/G	G/G	A/G	A/G
AM_10075	G/G	G/G	G/G	A/G	G/G	G/G	G/G	G/G	G/G

Start Preprocess

Figure 6. The input DMET SNP dataset represented as an interactive table. By selecting the subject to be assigned to class B and by clicking on the Start Preprocess button, the user can run the microPipe analysis.

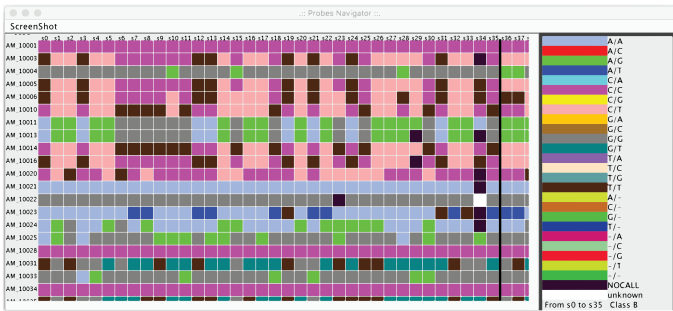


Figure 7. Visualization of the SNP distribution in each probe represented as a heat map.

DMET-DatasetPlusOS.txt :: DMET-Miner: Differences of ALLELES frequencies in classB with respect to classA ... /Users/giuseppagapito/Orapbox/RS45/datasets/D...

File	Analysis	About	Help
A	B	C	D
AM_10001	TA/TORS...	TA/TORS...	TA/TORS...
AM_10003	0.0	0.0	0.0
AM_10004	0.0	0.0	0.0
AM_10005	0.0	0.0	0.0
AM_10006	0.0	0.0	0.0
AM_10010	0.0	0.0	0.0
AM_10011	0.0	0.0	0.0
AM_10013	0.0	0.0	0.0
AM_10014	0.0	0.0	0.0
AM_10016	0.0	0.0	0.0
AM_10020	0.0	0.0	0.0
AM_10021	0.0	0.0	0.0
AM_10022	0.0	0.0	0.0
AM_10023	0.0	0.0	0.0
AM_10024	0.0	0.0	0.0
AM_10025	0.0	0.0	0.0
AM_10028	0.0	0.0	0.0
AM_10031	0.0	0.0	0.0
AM_10033	0.0	0.0	0.0
AM_10034	0.0	0.0	0.0
AM_10035	0.0	0.0	0.0
AM_10047	0.0	0.0	0.0
AM_10048	0.0	0.0	0.0
AM_10052	0.0	0.0	0.0
AM_10053	0.0	0.0	0.0
AM_10054	0.0	0.0	0.0
AM_10056	0.0	0.0	0.0
AM_10059	0.0	0.0	0.0
AM_10064	0.0	0.0	0.0
AM_10070	0.0	0.0	0.0
AM_10075	0.0	0.0	0.0

Correctors

Choose to activate one Correction:

☒ Without Correction

☐ Activate Bonferroni Correction

☐ Activate FDR Correction

Test significance: 95%

OK

Figure 8. Preprocessed table, from which user can select the statistical analysis to be performed. In the right corner, it is depicted the Statistical corrector panel, from which users can decide to apply or not a statistical correction.

As the final step, the mined rules are splitted and presented in two separate windows and ranked by confidence value, as conveyed in Figure 13.

The rules mined by microPipe, allow users to figure out the multiple relations among SNPs that are responsible for a particular response to a drug for one class in a pharmacogenomics study.

In summary, microPipe is a useful tool to conduct multiple analysis through a few mouse clicks.

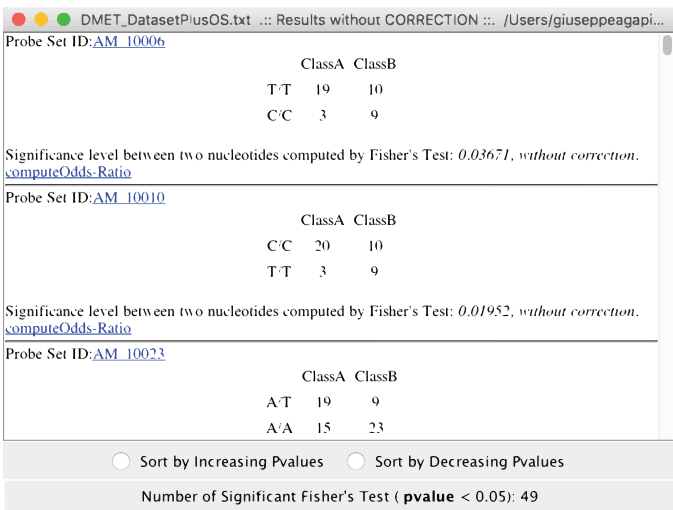


Figure 9. The relevant SNPs detected from microPipe by using Fisher’s Test.



Figure 10. Association Rule Mining warning message. microPipe advises the users that the mining phase is starting up.

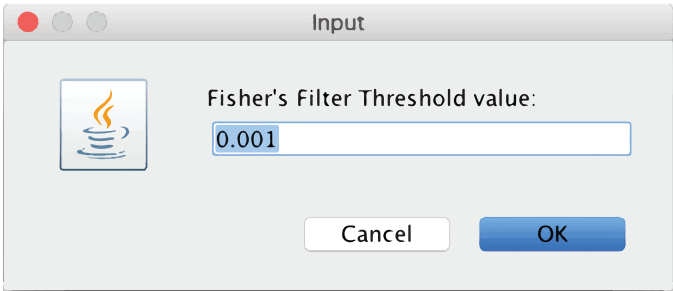


Figure 11. Window from which users can define the FisherFilter value with which to remove all the useless rows below this value.

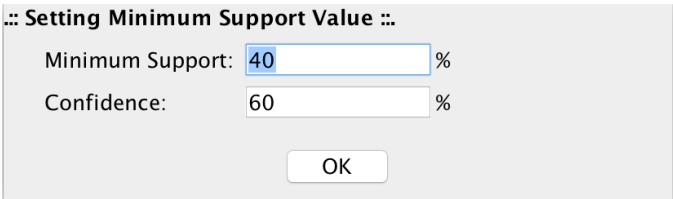
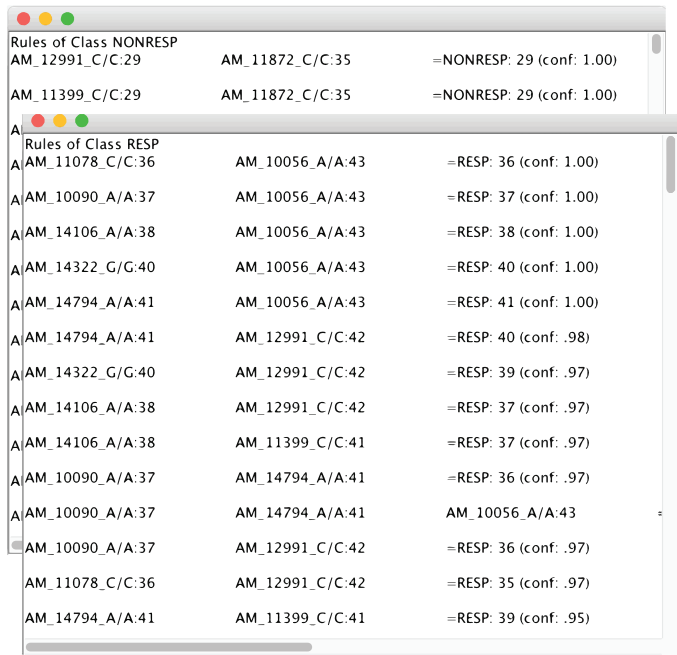


Figure 12. Minimum Support and Confidence setting up panel. In this panel, users can set the values of Minimum Support and Confidence with which microPipe computes the strong association rules.

It make possible to analyze an entire DMET SNP dataset with statistical and data mining techniques only by uploading the file to be examined.



Rules of Class NONRESP		
AM_12991_C/C:29	AM_11872_C/C:35	=NONRESP: 29 (conf: 1.00)
AM_11399_C/C:29	AM_11872_C/C:35	=NONRESP: 29 (conf: 1.00)

Rules of Class RESP		
AM_11078_C/C:36	AM_10056_A/A:43	=RESP: 36 (conf: 1.00)
AM_10090_A/A:37	AM_10056_A/A:43	=RESP: 37 (conf: 1.00)
AM_14106_A/A:38	AM_10056_A/A:43	=RESP: 38 (conf: 1.00)
AM_14322_G/G:40	AM_10056_A/A:43	=RESP: 40 (conf: 1.00)
AM_14794_A/A:41	AM_10056_A/A:43	=RESP: 41 (conf: 1.00)
AM_14794_A/A:41	AM_12991_C/C:42	=RESP: 40 (conf: .98)
AM_14322_G/G:40	AM_12991_C/C:42	=RESP: 39 (conf: .97)
AM_14106_A/A:38	AM_12991_C/C:42	=RESP: 37 (conf: .97)
AM_14106_A/A:38	AM_11399_C/C:41	=RESP: 37 (conf: .97)
AM_10090_A/A:37	AM_14794_A/A:41	=RESP: 36 (conf: .97)
AM_10090_A/A:37	AM_14794_A/A:41	AM_10056_A/A:43
AM_10090_A/A:37	AM_12991_C/C:42	=RESP: 36 (conf: .97)
AM_11078_C/C:36	AM_12991_C/C:42	=RESP: 35 (conf: .97)
AM_14794_A/A:41	AM_11399_C/C:41	=RESP: 39 (conf: .95)

Figure 13. Association rules mined by microPipe and conveyed in two windows related to the two analyzed classes RESP and NoRESP.

Indeed, this methodology frees researchers from having to spend time in repetitive steps such as adapting the dataset, preprocessing data, and so on that are tedious and error-prone tasks. Alternatively, by using microPipe, researchers can devote themselves exclusively to the analysis of results and their interpretation, in a much shorter time.

5. Conclusions

This paper presented microPipe a software pipeline that allows users to perform different analysis on DMET data, such as data mining, statistical and survival analysis through some mouse clicks. The easiness of microPipe makes it possible to speed up the investigation of DMET SNP data by some order of magnitude, because users are not involved in tedious and error prone tasks. The main strength of microPipe is the advanced computational engine that allows to perform several data analysis in parallel, by taking advantage of the available hardware. microPipe’s engine is designed to exploit the multi-CPU’s multi-Cores architectures to speed-up the computational analysis. The primary goal of microPipe is to support researchers in their daily activities, allowing them to analyze substantial heterogeneous datasets through several different kinds of analysis efficiently, without the need to manually check if the data are compatible with the chosen tools. The manual manipulation of a dataset is a tedious as well as an error-prone process, that could introduce bias into the data and consequently could produce poor accurate results. Conversely, by using microPipe, all the repetitive phases are done automatically by the tool. Thus the researchers can focus only on the analysis and evaluation of results, speeding up the process of knowledge extraction.

Acknowledgments: This work has been partially supported by the Data Analytics Research Center of the University Magna Græcia of Catanzaro.

Author Contributions: G.A. and M.C. conceived and designed the tool; G.A. implemented the tool; G.A. performed the experiments; P.H.G. and G.A. analyzed the data; G.A., P.H.G. and M.C. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SNP	Single Nucleotide Polymorphism
DMET	Drug Metabolism Enzymes and Transporters
ADME	Adsorption, Distribution, Metabolism and Excretion
NGS	Next Generation Sequencing
GWAS	Genome-Wide Association Studies
MS	Mass Spectrometry
OS	Overall Survival
PFS	Progression Free Survival
JDK	Java Development Kit
jar	Java Archive
FDR	False Discovery Rate
GUI	Graphical User Interface

References

1. Kuipers, O.P. Genomics for food biotechnology: prospects of the use of high-throughput technologies for the improvement of food microorganisms. *Current opinion in biotechnology* **1999**, *10*, 511–516.
2. Gupta, P.; Rustgi, S.; Mir, R. Array-based high-throughput DNA markers for crop improvement. *Heredity* **2008**, *101*, 5.
3. Ohnishi, Y.; Tanaka, T.; Ozaki, K.; Yamada, R.; Suzuki, H.; Nakamura, Y. A high-throughput SNP typing system for genome-wide association studies. *Journal of human genetics* **2001**, *46*, 471.
4. Arbitrio, M.; Di Martino, M.T.; Barbieri, V.; Agapito, G.; Guzzi, P.H.; Botta, C.; Iuliano, E.; Scionti, F.; Altomare, E.; Codispoti, S.; Conforti, S.; Cannataro, M.; Tassone, P.; Tagliaferri, P. Identification of polymorphic variants associated with erlotinib-related skin toxicity in advanced non-small cell lung cancer patients by DMET microarray analysis. *Cancer Chemotherapy and Pharmacology* **2016**, *77*, 205–209.
5. Di Martino, M.T.; Arbitrio, M.; Guzzi, P.H.; Leone, E.; Baudi, F.; Piro, E.; Prantera, T.; Cucinotto, I.; Calimeri, T.; Rossi, M.; Veltri, P.; Cannataro, M.; Tagliaferri, P.; Tassone, P. A peroxisome proliferator-activated receptor gamma (PPARG) polymorphism is associated with zoledronic acid-related osteonecrosis of the jaw in multiple myeloma patients: analysis by DMET microarray profiling. *British Journal of Haematology* **2011**, *154*, 529–533.
6. Martino, M.T.D.; Arbitrio, M.; Leone, E.; Guzzi, P.H.; Rotundo, M.S.; Ciliberto, D.; Tomaino, V.; Fabiani, F.; Talarico, D.; Sperlongano, P.; Doldo, P.; Cannataro, M.; Caraglia, M.; Tassone, P.; Tagliaferri, P. Single nucleotide polymorphisms of ABCG5 and ABCG1 transporter genes correlate to irinotecan-associated gastrointestinal toxicity in colorectal cancer patients: A DMET microarray profiling study. *Cancer Biology & Therapy* **2011**, *12*, 780–787, [<https://doi.org/10.4161/cbt.12.9.17781>]. PMID: 21892003.
7. Arbitrio, M.; Martino, M.T.D.; Scionti, F.; Agapito, G.; Guzzi, P.H.; Cannataro, M.; Tassone, P.; Tagliaferri, P. DMETTM (Drug Metabolism Enzymes and Transporters): a pharmacogenomic platform for precision medicine. *Oncotarget* **2016**, *7*, 54028–54050.
8. Guzzi, P.H.; Agapito, G.; Di Martino, M.T.; Arbitrio, M.; Tassone, P.; Tagliaferri, P.; Cannataro, M. DMET-Analyzer: automatic analysis of Affymetrix DMET Data. *BMC Bioinformatics* **2012**, *13*, 258.
9. Agapito, G.; Guzzi, P.H.; Cannataro, M. DMET-Miner. *J. of Biomedical Informatics* **2015**, *56*, 273–283.
10. Agapito, G.; Botta, C.; Guzzi, P.H.; Arbitrio, M.; Di Martino, M.T.; Tassone, P.; Tagliaferri, P.; Cannataro, M. OSAnalyzer: A Bioinformatics Tool for the Analysis of Gene Polymorphisms Enriched with Clinical Outcomes. *Microarrays* **2016**, *5*.
11. Guzzi, P.H.; Agapito, G.; Cannataro, M. coreSNP: Parallel Processing of Microarray Data. *IEEE Transactions on Computers* **2014**, *63*, 2961–2974.
12. Agapito, G.; Guzzi, P.H.; Cannataro, M. Parallel extraction of association rules from genomics data. *Applied Mathematics and Computation* **2017**.
13. Agapito, G.; Cannataro, M.; Guzzi, P.H.; Marozzo, F.; Talia, D.; Trunfio, P. Cloud4SNP: Distributed Analysis of SNP Microarray Data on the Cloud. Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics; ACM: New York, NY, USA, 2013; BCB'13, pp. 468:468–468:475.

14. A cloud framework for big data analytics workflows on azure; IOS Press, 2013; chapter Big Data.
15. Marozzo, F.; Talia, D.; Trunfio, P. Using clouds for scalable knowledge discovery applications. European Conference on Parallel Processing. Springer, 2012, pp. 220–227.