The usage of transcriptomics datasets as sources of Real-World Data for clinical trialling

Pedro Matos-Filipe[1,2], Juan Manuel García-Illarramendi[1,3], Guillem Jorba[1,2], Baldo Oliva[2*], Judith Farrés[1], José Manuel Mas[1*]

Affiliations and Contact Information

[1]Anaxomics Biotech SL, Barcelona, Spain;

[2]Structural Bioinformatics Group, Research Programme on Biomedical Informatics, Department of Experimental and Health Science, Universitat Pompeu Fabra, Barcelona, Spain;

[3]Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Barcelona, Spain

Corresponding authors:

* baldo.oliva@upf.edu (BO); xouse@anaxomics.com (JMM)

ABSTRACT

Background:

Randomized clinical trials (RCT) are limited in reflecting observable results out of controlled settings, which requires the execution of further lengthy observational studies. The usage of real-world data (RWD) has been recently considered to be a viable alternative to overcome these issues and complement certain clinical conclusions. Transcriptomics and other high-throughput data contain a molecular description of medical conditions and disease states. When linked to RWD, including demographical information, transcriptomics data is capable of elucidating nuances in disease pathways in specific patient populations. This work focuses on the construction of a patient repository database with clinical information resulting from the integration of publicly available transcriptomics datasets.

Results:

Patient data were integrated into the patient repository by using a novel post-processing technique allowing for the usage of samples originating from different/multiple Gene Expression Omnibus (GEO) datasets. RWD was mined from GEO samples' metadata and a clinical and demographical characterization of the database was obtained. Our post-processing technique, which we have named MACAROON – *MicroArray Cross-plAtfoRm pOst-prOcessiNg*, aims to uniformize and integrate transcriptomics data (considering batch effects and possible processing-originated artefacts). This process was able to better reproduce the down streaming biological conclusions in a 10% improvement when compared to other methods available. RWD mining was done through a manually curated synonym dictionary allowing for the correct assignment (95.33% median accuracy; 80.14% average) of medical conditions.

Conclusions:

Our strategy produced an RWD repository, which includes molecular information as well as clinical and demographical RWD. The exploration of these data facilitates the discovery of clinical outcomes and pathways specific to predetermined populations of patients by integrating multiple public datasets.

Keywords

RWD; Data Integration; Big Data; Microarrays; Post-processing; Normalization

BACKGROUND

Developments in the fields of Artificial Intelligence (AI) and Machine Learning (ML) (and their association with the pharmaceutical industry) have helped to boost/accelerate the timeline of production of pharmaceutical products, leading to a reduction in cost[1]. Techniques in the field are already being streamlined into production under various stages of the life cycle of pharmaceutical products[2], namely, drug discovery[3–5], pharmaceutical development[6], and clinical trial development and monitoring[7, 8]. Nonetheless, regulatory agencies still mostly admit data acquired through Randomized clinical trials (RCT). These institutions encourage the use of stored information about patients and molecular processes to reduce the number of RCTs whenever possible. When molecular data are available, it is pivotal to determine the relationship between the known clinical phenotypes and the available molecular patient description. If available, meaningful clinical patterns can be extracted from the molecular information. Obtaining molecular information in RCT can pose an obstacle from a practical and ethical standpoint since it can be very hard to find patients to conduct trials and obtain the data (i.e., orphan diseases), or it can be very hard to recruit patients without a specific goal from a deontological perspective. Furthermore, RCTs do not always reflect observable results out of controlled settings[9], which results in the execution of further lengthy observational studies. The usage of Real-world Data (RWD) has been recently considered to be a viable alternative to complement clinical conclusions. Despite some divergences (i.e., specific data sources), RWD is defined as data relating to patients' health status that did not originate in traditional clinical trials. Since 2018, both the US Food and Drug Administration (FDA)[10] and the European Medicines Agency (EMA)[11] have laid down their regulatory frameworks from which the usage of RWD could be used to obtain valid clinical evidence. In both cases, the accuracy and reliability of RWD (defined according to the scope of the study[12]) were clear requirements for its acceptance for regulatory purposes[11, 13].

In 1970, Francis Crick presented the Central Dogma of Biology[14]. According to his theory, information that codes proteins in living organisms flows from the genome, establishing the relationship between the individuals' genome and the phenotype it would display. Although the current paradigm now includes a more dynamic approach in comparison to the one Crick presented, the main idea that genetic material influences the proteome and, consequently, the remaining molecules contained within the organism, remains unrefuted[15]. Therefore, the analysis of an organism's genome and transcriptome sheds light on molecular events underlying human biology and disease[16], producing significant insight into its phenotype, and giving key clues into how to influence it.

The advent of high-throughput (HT) biological techniques in the 1990's made available the analysis of multiple targets at once, from multiple perspectives (e.g., proteomics, metabolomics, genomics). HT techniques like microarray (MA) (1997)[17] and RNA Sequencing (RNA-Seq) (2006)[18] allow the characterization of the samples' transcriptome in a quick way that is light on labour and costs. When compared to MA, RNA-Seq presents some advantages concerning sensitivity[19] and dynamic range[20][1] and is particularly suited for *de novo* sequencing since it does not require reference transcripts for probe design[21]. On the other hand, MA shows higher throughput, is less labour intensive[22] and, due to being an older technology, presents more currently available data[23] which is a key aspect when dealing with big data studies[24, 25]. Gene expression values seem to correlate in both techniques[20, 26, 27], which validates the usage of any of them if limitations are considered.

The usage of AI, and specifically ML algorithms, is dependent on high volumes of data[28]. The great volumes of data available today led to the adoption of a culture of sharing and reusing HT-generated information within the scientific community, and to the development of public

[1] The concept of dynamic range refers to the ratio between the maximum and minimum value a specific technology or device may withstand.

biological databases to store them. After 30 years, the amount of data hosted yearly in these repositories reaches Terabytes, and it is expected to reach Exabytes by 2025 – surpassing data created by social media platforms[29]. **Table 1** lists four databases containing transcriptomics experimental data. The Gene Expression Omnibus (GEO)[30] is, to our knowledge, the database with the highest number of available samples and experiments. However, a high number of entries on the platform only include data already pre-processed by the submitting author. A large variety of methods to background correct, normalize and summarize transcriptomics data[31, 32] are being used, with no consensus. To avoid this issue, most current protocols for data integration involve the usage of raw data[33] and the choice of samples from the same experiment[34]. As last resort, the usage of samples involving the same techniques, machines and processing methods is also commonly used (e.g., GEO datasets[35]). Consequently, a large portion of data included in these transcriptomics repositories is usually excluded from data integration methodologies and protocols. Furthermore, the metadata accompanying the processed files often consists of inaccurate information regarding the processing procedures which completely invalidates its usage; or contains too complex vocabulary[36, 37], grammar and sentences which blocks its usage in experiments relying on Big Data given their lack of machine interpretability[38]. Nonetheless, if properly processed and transformed, biological datasets mined from GEO can serve as RWD sources. By exploring the association between molecular information (harvested from gene expression values) and clinical variables (obtained from metadata reported in each sample), it is possible to obtain accurate evidence to support a clinical outcome upon a given treatment.

Table 1. Biggest genome databases and the number of available human samples and experiments (data obtained from each platform on the 3rd of June 2021). * "NA" indicates that the data was not available on the platform.

Database	**Number of Samples**	**Number of Experiments**	**Ref**
Gene Expression Omnibus (GEO)	2,293,223	66,569	[30]
ArrayExpress	1,309,771	27,068	[39]
Japanese Genotype-phenotype Archive (JGA)	278,601	NA	[40]

The European Genome-phenome Archive (EGA)	NA	4,432	[41]

Lately, several scientific works were published exploring the usage of transcriptomics datasets as RWD sources. In 2018, Arisi et al.[42] discovered a subset of comorbidities that help discriminate diagnostic subgroups of Mild Cognitive Impairment by analysing transcriptomic datasets containing gene expression information from American and European cohorts. More recently, Kato et al.[43] linked HT transcriptomics data associated with clinical information to propose a set of molecular variables to help select chemical treatments for cancer patients. A board of biomarkers was tested against a set of patients. Those who were treated according to their biomarkers' levels achieved significantly better clinical outcomes, including longer progression-free and overall survival. Fernandes et al.[44], on the other hand, extracted possible breast cancer biomarkers by linking RWD data with RNA Seq data. The analysis allowed the subtyping of patients and the prediction of clinical outcomes accordingly.

In this paper, we provide a global outlook of the structure of GEO MA studies as sources to obtain data-driven clinical evidence. Firstly, given the heterogeneity found in HT data, we developed a novel technique which corrects pre-processed MA data obtained from multiple experiments and different processing techniques that we have named MACAROON – *MicroArray Cross-plAtfoRm pOst-prOcessiNg*. First, a text-mining process was applied to standardize the existing clinical labels in the HT repository and to identify exclusively patient data (excluding non-human data and data coming from immortalised cell lines). MACAROON uses language processing techniques to build machine learning models that highlight the transformations applied at the original submission, overcoming obstacles related to poor metadata. Then, we integrated a protocol that uses House Keeping (HK) genes' expression values to readjust the previous different normalization techniques – corresponding to the normalization methods used in MA datasets before upload to GEO. By applying the

methodology herein proposed, the researcher is able to combine MA profiles from multiple sources or experiments independently of the samples' pre-processing steps. On the other hand, we show how text mining can standardize metadata, namely by classifying available samples according to health status and demographics – thus building a source of RWD. The usage of GEO can link these RWD with molecular motifs by exploring the transcriptomic patterns underlying the provided HT experiments. With this effort, we aim at easing access to RWD, which is currently the main limitation of pharmaceutical companies in this kind of studies[45].

RESULTS AND DISCUSSION

The foremost objective of this work is to build an RWD source by using the molecular and phenotypical data stored in the GEO database[30]. We have obtained a total of 1,866,292 MA samples (including gene expression and corresponding metadata). Gene expression data were subjected to the novel post-processing method MACAROON (introduced in this work), while clinical metadata was mined for phenotypical labels (demographics and medical conditions).

Clinical labelling in GEO is often restricted to medical conditions of interest to the specific experiment to which they belong. Consequently, further medical conditions may be disregarded, even in cases in which these are highly probable. For instance, patients with ageing-associated conditions usually display other disorders of the same type. As such, phenotypical labels in combination with the corresponding gene expression data were used to generate classifiers aimed at predicting clinical phenotypes excluded from metadata. The whole set of samples was run through these classifiers and, for each sample, newly predicted clinical labels were generated – further complementing our RWD repository.

In summary, each sample in the RWD repository consists of the post-processed MA data, and the corresponding clinical and demographical labels. To better guide the reader through the methods herein presented, an overall depiction of the pipeline deployed in this work is

schematized in **Figure 1**. This new resource links the genotype of the patient (collected from the molecular description included in the transcriptomics data – **Figure 1A**) to its phenotype (gathered from the clinical and demographical data – **Figure 1B**). Phenotypical data includes information mined from GEO files (figure 1B.1) and complemented with ML predictions of further clinical labels (figure 1B.2).

Figure 1. The overall pipeline of the methods applied in this work. A. Collection and processing of MA data. B. Collection and processing of RWD from GEO.

Global microarray data processing

MACAROON is split into two stages. First, we revert and clean the processing procedures applied by the original researcher (i.e., reversal of logarithm procedures and labelling of non-intensity files) to uniformize the orders of magnitude in the samples' data. Then, we apply a GEO-wide normalization process to reduce the impact of sample-specific background corrections – which could lead to misleading biological conclusions.

ML-based assignation of samples' pre-processing

Metadata accompanying the MA samples were downloaded for all GEO samples. Each text file was analysed in order to label samples for the presence of terms related to the application of different logarithms, standard MA normalization methodologies and types of data (i.e., intensity or differential expression (DE) results) – that we will describe as "pre-processing" in the rest of this paper. A combination of interest terms was used to select samples in datasets to train ML classifier models – that we will call "$ML_{processing}$" moving forward. These terms were picked to maximize the distinguishability of "applicable" (when a processing method was applied) from "non-applicable" classes (when that specific method was not applied) – see Methods section for details (**Table 4**). This process resulted in five different MA datasets, each with 5000 samples (70-30 train/test split): application of (1) Log2, (2) Log10, (3) Log

(unknown base), (4) DE ratios and (5) Normalization. We then used these datasets for training $ML_{processing}$ for inferring which pre-processing methods were applied to MA samples whose metadata was inconsistent or non-existent.

The datasets mentioned above were used to train $ML_{processing}$ to infer if each determined processing technique was applied to the sample or not. From each sample in the dataset, we extracted a statistical description (minimum, maximum, mean, standard deviation, and variance) and the values of 5 low-variability HK genes (top five genes in **Table 3** in the Methods section). The resulting best-performing models were based on decision trees and Support Vector Machines (SVM) algorithms. Given that training datasets were built with class balancing in mind, we analysed models' results based on accuracy (ACC), corresponding to the proportion of correct predictions provided by the models (Equation 1). In our results, models could attain a satisfactory level of accuracy (25-k fold cross-validation (CV); in the range [70%-98%]). To maintain a maximum certainty in all predictions, we explored the accuracy of the top-three models in each 5%-quantile of the score range in that model. We found that Log2, Log10, Log and Ratio were capable of outputting statistically significant classifications (95% correct answers per score quantile) in both classes (see **Supplementary Figure 2**). Yet, in Normalization, ML models did not display satisfactory prediction capabilities due to difficulties in clearly selecting samples on the negative class, which made us disregard them in further steps. Given the large number of missing values in the training datasets (especially concerning HK MA intensities), the various models were trained with a different set of features, as a result of an automatic feature selection strategy. Final models for each processing procedure being evaluated comprise an ensemble of the top-three best performing ML models in each category (which allowed us to avoid missing value imputation approaches). Exceptions reside in models predicting Log2 and Log10, where ensembles also contain models to distinguish if Log with an unknown base (log_x) was applied (a total of six

models). In **Table 2**, we display the details and CV performance metrics of each model included in the three ensembles. This process consists in dividing the dataset into k folds (in our case 25-fold). The $ML_{processing}$ are, then, trained sequentially with k-1 parts and tested using the remaining set[46]. CV metrics consist of the means of the metrics obtained in each training procedure. Here, three further metrics are shown: (1) precision (PRE), corresponding to the proportion of positive results that are true positive results (Eq. 2); (2) sensitivity (SNS), corresponding to the proportion of positives that are correctly identified (Eq. 3); and (3) specificity (SPC), corresponding to the proportion of negatives that are correctly identified (Eq. 4). We assessed the performance of the various ensembles considering the training dataset, the testing dataset (both following the requirements displayed in Table 4), and against all samples in the dataset. Given they were carefully defined, metrics corresponding to the train and test sets present very similar results. However, metrics computed using all the samples available present a significant drop across all metrics (average 44.39%) except for SNS (where differences between train/test datasets and all samples average 0,30% across the various ensembles). We believe that this effect relies on the poor definition of negative classes. Given that the text mining process works by identifying specific words present in the metadata data of the various samples, negative classes may be poorly assigned if the data was pre-processed using a technique that was not mentioned in the corresponding metadata. This problem is found frequently in the literature[34, 36–38]. Further test performance metrics are available in **Supplementary Table 1** (training metrics can be found in **Supplementary Table 2**). A comparable assessment of these results can be found in **Figure 2**.

$$AAC = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{(Eq. 1)}$$

$$PRE = \frac{TP}{TP + FP} \quad \text{(Eq. 2)}$$

$$SNS = \frac{TN}{FP + FN} \quad \text{(Eq. 3)}$$

$$SPC = \frac{TN}{FP + TN} \quad \text{(Eq. 4)}$$

*where TP corresponds to True Positives; TN, to True Negatives; FP, to False Positives; and FN, to False Negatives.

Table 2. Composition of ensemble models for the classification of previous processing.

Dataset	**ML Algorithm**	**Included Features**	**Feature Selection Method**	**CV-ACC**	**CV-PRE**	**CV-SNS**	**CV-SPC**
Log2	SVM	MIN, MAX, AVG, STD, VAR, Q99497, Q15843, P28070, P84090, P28074	Elastic Net	94,31	91,40	96,86	92,10
	SVM	MIN, MAX, AVG, STD, VAR	Simple Regression	92,99	94,24	91,92	94,11
	SVM	MIN, MAX, AVG, STD	Random Forest	92,53	93,60	91,67	93,43
Log10	SVM	MIN, MAX, AVG, STD, VAR, Q99497, Q15843, P28070, P84090, P28074	Elastic Net	94,48	96,56	92,88	96,28
	SVM	MIN, MAX, STD, VAR	Simple Regression	92,28	91,05	93,23	91,39
	SVM	MIN, MAX, AVG, STD, VAR	Wilcoxon	92,10	91,14	92,70	91,54
Log	TREE	MIN, MAX, AVG, STD, VAR, Q99497, Q15843, P28070, P84090, P28074	Elastic Net	98,36	97,98	98,08	98,56
	TREE	MIN, MAX, AVG, STD, VAR	Ridge	96,77	96,94	96,89	96,64
	TREE	MIN, MAX, AVG, STD	Random Forest	96,47	96,54	96,73	96,19
Ratio	TREE	MIN, MAX, AVG, STD, VAR	Entropy, Correlation	81,13	81,00	81,48	80,78
	SVM	MIN, STD	None	78,82	71,50	96,05	61,50
	SVM	MIN	Elastic Net	78,39	70,70	97,20	59,49

Figure 2. Global test metrics of the various ensemble models in $ML_{processing}$. (A) Log2, (B) Log10, and (C) Ratio

After training of $ML_{processing}$, samples where pre-processing metadata was unclear were classified using these models. Classification results were used to revert all samples to the original state by reverting the specific Log operation (see Methods section for detail). Instead of uniformizing the data by applying Log, we decided to revert the Log transformation in all samples (see **Pseudo-code 1**). Such decision allowed us to take the mathematical nature of the Log operation into account. Given that Log(x) does not exist if x is negative, to avoid approximation of imaginary numbers, values were restored to their original form.

Pseudo-code 1. Composition of ensemble models for the classification of previous processing.

```
for MA sample in GEO
    if log processing is known
        log processing = log processing
    else
        log processing = MLprocessing(raw MA sample)
    end if

    if is raw is True
        do nothing
    else if log processing is log2 and not log10
        uniform MA sample = 2^(raw MA sample)
    else if log processing is log10 and not log2
        uniform MA sample = 10^(raw MA sample)
    else
        do nothing
    end if
end for
```

House Keeping-based regression for normalization

After the first stage of this method – aimed at cleaning and uniformizing original processing techniques –, we designed a normalization method based on HK gene expression. A set of 40 HK genes is built using their median expression value with the least cross-tissue variability as found in the literature. In each sample of the whole dataset, genes were mapped in a two-coordinate system and a regression line was fitted using a linear (polynomial regression with one coefficient) and a non-linear (spline regression) method. Although MA normalization protocols reliant on regression techniques usually use the LOWESS algorithm[47], we were prevented from using such a strategy given that it does not allow for extrapolation. Given that HK genes are rarely under/overexpressed, extrapolation was a requirement in order to account for under/overexpressed genes.

For benchmarking, a test dataset containing MA data from prostate cancer samples and healthy controls (available in **Supplementary Table 3**) was extracted. Their expression values were then visually explored according to their distribution sample- and gene-wise; representations of these results are available in **Figure 3**. This figure shows a progressive representation of the datasets as the stages of MACAROON are applied. Taking the general assumption that most

genes should not see any change of expression across samples[48], the mean intensity should show a tendency towards a constant value. This tendency is visible from the raw datasets (F3A) to the data after previous pre-processing correction (F3B), and then to the data subjected to HK regression (F3C and F3D). The effect is also clear on the MA-plots (c – bottom of Figure 3), where M is defined by equation 5 and A is defined by equation 6. Assuming that the expression of most genes stays generally constant across samples, the M value corresponding to the great majority of genes should stay constant along with A if the data is well normalized. The red line in F3c corresponds to a regression line of the M/A pairs in the plot. If the initial assumption is met, the red and blue lines (corresponding to $M = 0$) should overlap. This is the case for each step of Trans Babel, where the blue and red lines show a tendency to overlap and better depict a biological scenario. On the other hand, samples present a strong localized concentration of data points (see density curves in Figures 3A and B). We attribute this event to the lack of dynamic range that is characteristic of MA[20].

$$M = log_2\left(\frac{Test_i}{Control_i}\right) \qquad \text{(Eq. 5)} \qquad A = \frac{1}{2}log_2(Test_i \times Control_i) \qquad \text{(Eq. 6)}$$

*where $Test_i$ and $Control_i$ correspond to the mean intensity at the gene i.

Figure 3. Log-expression data distribution across the various stages of MACAROON. A, B, C and D refer to the datasets produced along the process: A. Raw unprocessed data; B. After reverting preprocessing; C. After spline regression; and D. After polynomial regression. (a), (b), and (c) concern the samples displayed in the figure: (a). healthy control samples; (b). prostate cancer samples; and (c). MA-plot (control vs. prostate cancer) – as described in Eq. 5 and 6.

MACAROON compared against the YuGene method

As further validation of MACAROON, we compared its output against YuGene[49]: to our knowledge, the only available method that can also post-process MA data (without depending on the availability of raw data). **Figures 4A, B and C** present a sequence of bar plots describing

the intensity values of the GEO series GSE102124[50] (one of the samples available at the PCa dataset used above in this work). The sub-plots correspond to these data in the preprocessed form, followed by the application of MACAROON and YuGene. From the analysis of the figure, we can understand how each one of the methods transforms the raw data and how they impact the final biological interpretability. One of the major drawbacks of MA, in comparison to RNA-Seq, is its low dynamic range. In MACAROON, an effort is made to maintain the standard range of processed MA. On the other hand, in YuGene, values are scaled into a 0-1 range. Given that the output from both methods is equivalent to the log2 scale, the application of MACAROON provides better compatibility with current methods when performing DE analysis – this fact was later observed when performing a comparative DE analysis with the three datasets. However, MACAROON shows worst performance than YuGene at aligning the samples' means, which also leads to a change in the range of DE results. Different MA platforms have different coverage, a concept that portrays the probes that are included and the genes that are assessed by using each platform. Consequently, datasets obtained from the integration of different platforms leads to datasets with many missing values. Our normalization approach, based on regression, allows the conservation of maximum and minimum values per probe, while supporting missing values, and without imputing them; retaining the original biological motifs in the dataset.

Figure 4. Direct comparison between methods. In the left column, we show the intensity per sample. A corresponds to the log2 raw data; B, using MACAROON; and C, using YuGene[49]. In the right column, we display a comparison of the results of a Differential Expression analysis. In D, we show a dispersion of the log Fold Change values; in E a Venn diagram portraying common differentially expressed genes depending on the dataset used.

As means of further comparing the three datasets, we ran a DE Analysis with data processed using the different methods. In **Figure 5D**, we represent the results' space (density of log Fold-change - FC) of the DE according to the dataset. Given the original range of the dataset, none

of the methods originated logFC values above 1. Additionally, and as discussed previously, MACAROON and YuGene tend to provide results in a reduced range (in comparison to the original dataset) because of the misalignment of samples' means or by the heavy scaling of datasets, respectively. Nonetheless, when looking at overlapping differentially expressed genes (**Figure 4E**), MACAROON shows greater similarity to the original results (13% matching) in comparison to YuGene (3%).

Case Study: Differential Analysis Experiment

For presenting the usability of this method in the real world and to test it by comparing possible biological conclusions, we have used the same dataset used for the metrics presented in the previous sections in yet another DE analysis experiment. We have used both spline and polynomial regressions strategies as means of assessing differences in biological interpretability. The overall results using each regression technique are displayed in **Figures 5A** and **B** via volcano plots, the full list of results is available in the Supplementary Material (**ST 4**). The analysis of these materials elucidates that both regression methods are equivalent, which suggests that spline regression detected a linear pattern in these MA data. This conclusion goes in line with descriptions in the literature. According to the characterization provided by David, E.[51], the intensity value detected by the MA scanner is equivalent to the actual value of expression multiplied by two constants (equation 7). Through the proposed mathematical equation, the error at probe and sample levels are considered in a linear relationship. Significant differences in the DE results provided by the two methods differed on only two genes: KRTAP20-2 (Keratin-associated protein 20-2) and MT-ND1 (NADH-ubiquinone oxidoreductase chain 1); where expression differences are significant in polynomial regression, but not in the spline case.

Figure 5. DE analysis results using spline and polynomial regressions. A, B. Volcano plots of the DE using both regression techniques. C. Enrichment analysis results using the polynomial dataset (where BPs found in other works in literature are highlighted in red). In A and B, the dots in red are under-expressed genes and the dots in

blue are over-expressed genes (prostate cancer vs. control). In C, colours represent the first rank of GO BPs, while each column refers to the second rank.

$$i_{ps} \sim k_s \cdot \alpha_p \cdot \theta_p \qquad \text{(Eq. 7)}$$

*where i_{ps} is the intensity value for probe p and sample s; k a constant value concerning the sample; α a constant value concerning the probe; and θ is the actual value of expression.

To compare our results with those found in the literature, we ran a gene enrichment experiment against Gene Ontology[52] Biological Processes (BP) using PANTHER[53]. Given the theoretical linear behaviour of MA data discussed above, we decided to take polynomial regression as the technique for further steps. Results are presented in **Figure 5C**. Here, BPs found in other publications on Prostate Cancer are highlighted in red. Yang *et al.*[54] reported differences in biological adhesion, while Fan *et al.*[55] described changes in cell cycle control, cell death regulation, cell development, and focal adhesion. Sun *et al.*[56] and Verma *et al.*[57] reported alterations in the hippo signalling pathway, negative regulation of growth, and multiple metabolic progression. However, the same author also stated a relatively decreased mRNA expression level of PTEN in patients with PTEN mutations, a result we did not confirm.

Generation of the Real-World Data repository

Mining of Demographic Features

The demographical characterisation of patient populations is of utmost importance in the context of clinical trials. The overall outcome of a clinical intervention may differ significantly depending on age[58, 59], biological sex[60, 61] or ethnical background [62, 63]. Additionally, in the case of drug studies, information about body weight is also crucial for the analysis/simulation of pharmacokinetics[64–66] and to better understand other unexpected effects (e.g. "BMI paradox"[67, 68] – in which, despite their otherwise higher risk level,

extreme weight patients seem to have better clinical outcomes in metabolic-related diseases). Consequently, we mined GEO for information on five clinical variables, namely age, body mass index (BMI), height, and ethnicity. Given the unstructured nature of GEO metadata, few samples are fully described according to these features, which required the creation of different datasets. Furthermore, demographic features were often embedded in natural text – which led us to apply natural language processing (NLP) techniques to extract and organize them. In **Figure 6**, a graphical description of the obtained GEO patient populations is presented. Visual analysis of the data suggests that populations seem to be skewed towards a Western representation of demographics. This fact was highlighted by the ethnological dispersion (figure 6D) of patients, supported by values consisting of age, BMI, and height (figure 6A, figure B, and 6C, respectively). However, the comparison against the reference demographics of these populations was not significant (results from these statistical tests can be found in **Supplementary Table 5**). As such, despite being able to simulate a pool of patients (following predetermined demographic distributions), we cannot extract prevalence data given that the global population of samples does not follow what we observe in reference to regional populations.

Figure 6. Demographic description of MA samples in GEO. A, Age distribution of samples (n=152,237). Colours represent various age groupings: light yellow – childhood; yellow – adulthood; orange – elderhood. B, BMI distribution of samples (n=14,243). Different shades of red represent different gradings of BMI: very light – underweight; light – normal; red – pre-obese; dark– obese class I; very dark – obese class II. C, Height distribution of samples (n=6,200). D, Ethnicity background ratios (n=33,465). E, Biological sex ratio of samples (n=188,618). F, MeSH terms of disease conditions (top leaf in the MeSH ontology tree) found in GEO samples. G, balanced accuracy of the $ML_{clinical}$ models aimed at predicting samples' medical conditions. Results are grouped by the MeSH family. Each black dot corresponds to the average balanced accuracy of a medical condition in that family.

Mining demographical metadata from GEO samples is not a new subject. Other works have accomplished this task either by manual curation (STARGEO[69], and OMiCC[70]) or by applying AI and NLP techniques (MetaSRA[71] and ALE[72]). However, these works

disregard some parts of the demographical description by exclusively reporting the age of the patient and, exclusively in the case of ALE, gender. Additionally, the methods applying manual curation are restricted to the samples presently included in their database, which excludes its usage in later works. Further differences among these tools were discussed by Wang et al. in their review of the subject[73]. In our dataset, we have gone a bit further than current methods in the literature by exploring a larger number of demographical features. Furthermore, our pipeline considers the various forms under which demographical information is conveyed in GEO. Our text-processing methodology standardizes synonyms and units (e.g., age values, that may come in weeks, months or years were converted to years in cases; an array of available gender notations were converted to "male" and "female") which allows full machine interpretability and integration.

Mining of Medical Conditions

A similar strategy was used to mine for information on medical conditions associated with the patient samples. The title of each GEO-GSE experiment study and the accompanying text (present in GEO samples' metadata) was used to assign mentioned medical conditions and their status as healthy, control, or intervention case. The Biological Effectors database (BED)[74] is a database that relates biological processes (adverse events of drugs, drug indications, medical conditions) with the proteins most closely associated with them, therefore making a molecular description of pathways in humans. It results from the manual parsing of literature to list possible genes and proteins associated with specific medical conditions, and the role they display (activation state and further molecules that influence or are influenced by a given protein or gene) in specific disease pathways. This strategic advantage results in the significant reduction of automation-related errors in the database in comparison with other databases on the subject (e. g., DisGeNET[75] or KEGG[76]). Extracted pathological statuses in GEO samples were mapped to the BED. This process was evaluated by manually checking if BED

terms were correctly assigned, attaining a median accuracy of 95.33% (detailed accuracy per medical condition is displayed in **Supplementary Table 6**). Linking these terms to BED permitted associating each medical condition and consequently each sample in GEO to molecular descriptions of disease. Figure 6F describes the overall distribution of GEO patients across Medical Subject Headings (MeSH)[77] conditions groups. As in the case of the demographical description, the distribution of diseases highlights the fact that patient samples in GEO reflect the interests of the scientific community and corresponding funding agencies. Most common medical conditions correspond to those mostly investigated, and not to those more common in societies (e.g., cancer, neurological and psychological diseases).

Prediction of Medical Conditions

Commonly, clinical annotation in GEO is restricted to medical conditions concerning the experiments' scope. To further complement medical condition annotation on the dataset, we resorted to ML classifiers – that we will refer as $ML_{clinical}$ moving forward. The input architecture of these models was defined by the BED medical condition it predicts. Features were defined by BED information on disease-related proteins. Specifically, genes/proteins annotated as being linked to the models' phenotype defined what information is taken as input by each $ML_{clinical}$ model. However, due to the different MA platform coverage displayed by the samples – a subject discussed in previous sections of this work, various models were trained by BED medical condition in order to be able to counteracting the large number of missing probes. A global depiction of the cross-validation balanced accuracy (BACC – Equation 8) metrics is displayed in Figure 6G. Here, given the large number of BED conditions (>250), different pathologies are grouped into MeSH condition groups – as in Figure 6F; for further detail, extensive results are available in **Supplementary Table 7**. Across all available BED diseases, $ML_{clinical}$ models reached at least 81.57% balanced accuracy (average: 91.47%; standard deviation: 0.044). All $ML_{clinical}$ models attaining over 80% CV-BACC were kept, and

sequentially used (sorted in descending order according to their BACC) to classify MA samples in the RWD repository. However, we were unable to produce viable $ML_{clinical}$ models for 59.02% of BED medical conditions mainly due to a lack of GEO samples from those specific pathologies. Regardless, the overall accuracy obtained through these models proves MACAROON's suitability for ML experiments (additionally to the DE analysis presented and discussed in previous sections). Additionally, these models will also be used for filling the gaps in missing metadata, as well as studying possible comorbidities.

$$BACC = \frac{SNS + SPC}{2} \qquad \text{(Eq. 8)}$$

CONCLUSIONS

In this work, we present a global pipeline for the usage of the GEO database as a source of RWD data. Our global goal was to link demographical and clinical information (derived from GEO samples metadata) with the patients' underlying molecular patterns (derived from MA data profiles).

Given that most data in GEO are available in its processed form, we start by presenting a method that ingrates MA data from multiple different sources (i.e., different MA platforms) onto a comprehensive MA dataset. Here, we focused on being capable of using samples from previous experiments when raw (unprocessed) intensity values are not available, and only processed and normalized files are provided by the investigators. The different processing techniques for the normalization of MA data, and the fact that often supplementary metadata lacks precise information, negatively influence the standardization process and the conduction of data integration for Big Data experiments. To overcome these challenges, our processing

protocol consists of two main steps: (1) samples metadata enrichment using a natural-language processing algorithm; and (2) HK-based normalization using cross-GEO background datasets.

We have tested each intermediary step independently, and the results show great improvement as we have gone through each stage. The usage of MACAROON was tested in a prostate cancer case study and could reproduce the biological conclusions found in previous works from the literature analysing the transcriptome of the study samples. These results support the usage of MACAROON in future transcriptomics experiments as a novel approach for MA data integration and the exploration biological patterns in pre-processed data. Moreover, it runs independently of previous data processing, allowing the usage of bigger and more diverse cohorts – as required for the discovery of new, patient-centred biomarkers and pharmacological targets. MACAROON has been demonstrated to be a good method for integrating data from different MA datasets when compared to other available methods – better depicting the biological motifs in the original pre-processed datasets.

Additionally, we explored the usage of GEO metadata as a source of RWD. In addition to the medical conditions linked to each sample, we took the example of RCTs and mined for demographical information. This information, besides describing the patient population, allows to assess possible clinical nuances dependent on age, biological sex, ethnicity, and body mass distribution. Despite not being able to significantly modulate any predefined human population, which for instance excludes the possibility of prevalence studies, our current data allow us to randomly choose samples belonging to a custom demographical distribution, according to the research purpose and interest. Expectedly, a larger portion of patients' metadata does not comprise these data. Future steps to this work would indeed include the usage of ML to extract this information from patterns within the MA data itself. Using these methodologies, it will be possible to include further samples in clinical experiments and explore an even more diverse set of patients.

METHODS

MACAROON, transcriptomics data obtention and processing

Obtention of GEO data and dataset generation

Transcriptomic datasets were constructed using MA data available in the GEO repository[30, 78]. We have accessed all the available MA human samples (as of 22/12/2020) using the provided "Repository browser" tool. Through this process, we obtained a total of 27,173 series. Each GEO series was divided into its different samples, providing 2,017,019 MA expression profiles corresponding to the column "Value" on the sample page. We kept all the data files in their processed form (as made available by the submitting researcher). Moreover, the metadata matrix (included in the "Data table header descriptions" table of the sample page) was also kept and used in subsequent methods. This step was performed with Python 3.9.0 using the package Beautiful Soup 4[79]. We obtained a total of 1,866,292 samples containing processed MA data files. 211,384 of these did not contain any expression data (were empty entries in the database).

As probe names are platform-dependent (different platforms use different ids for the same genes), probe ids were uniformized to reflect the corresponding gene. The soft file ("GPLXXX_family.soft.gz") corresponding to the platform mentioned in the samples' metadata was downloaded and processed. To account for "custom platforms" and maximize speed, a local cross-GEO database including platform-specific ids and the corresponding Entrez ID was constructed, and expression files were reindexed according to it. Genes that included multiple intensity measurements were summarized and their mean intensity was kept. To maximize compatibility with the BED, probes were also mapped to the corresponding Uniprot ID.

MA integration (postprocessing and normalization)

Composition of datasets

The MA datasets mentioned in the previous steps were statistically described, and information on the maximum value, minimum value, average, standard deviation, and variation was extracted. Additionally, the value corresponding to the expression of five HK genes was kept. HK genes are responsible for the basic functions of the cell[80] and are universally expressed in all tissues and cell types[81]. Given their known variation in the expression levels in some diseases or conditions[82], we increasingly ranked known HK genes according to the standard deviation of their expression across multiple tissues, as published by Eisenberg et al. (2013)[83]. Additionally, the ranked list was filtered to only include genes appearing in a minimum of 13 different previously available lists, as reported by Zhang et al. (2015)[84]. The final HK genes list is available in **Table 3**.

Table 3. List of HK genes included on the datasets used in this work organised by standard deviation of expression across multiple tissues. The top 5 genes of this list were used in the datasets for sample classification.

Uniprot	**Gene Name**	**Protein Function**	$\boldsymbol{\sigma(log_2(RPKM))}$
Q99497	PARK7	Parkinson's disease protein 7	0.273137
Q15843	NEDD8	NEDD8	0.350249
P28070	PSMB4	Proteasome subunit beta type-4	0.370264
P84090	ERH	Enhancer of rudimentary homolog	0.382595
P28074	PSMB5	Proteasome subunit beta type-5	0.397559
P19338	NCL	Nucleolin	0.423427
P78344	EIF4G2	Eukaryotic translation initiation factor 4 gamma 2	0.446228
P84077	ARF1	ADP-ribosylation factor 1	0.458529
P15954	COX7C	Cytochrome c oxidase subunit 7C, mitochondrial	0.474006
P49773	HINT1	Histidine triad nucleotide-binding protein 1	0.474979
P50395	GDI2	Rab GDP dissociation inhibitor beta	0.496972
P00441	SOD1	Superoxide dismutase [Cu-Zn]	0.502851
P37108	SRP14	Signal recognition particle 14 kDa protein	0.504101
P61803	DAD1	Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit DAD1	0.519353
P18583	SON	Protein SON	0.536725
P0CG47	UBB	Polyubiquitin-B	0.537386

P00558	PGK1	Phosphoglycerate kinase 1	0.547633
Q06830	PRDX1	Peroxiredoxin-1	0.549351
P48556	PSMD8	26S proteasome non-ATPase regulatory subunit 8	0.552184
Q92572	AP3S1	AP-3 complex subunit sigma-1	0.581663
P09429	HMGB1	High mobility group protein B1	0.595868
P25705	ATP5F1	ATP synthase subunit alpha, mitochondrial	0.60272
P48201	ATP5A1	ATP synthase F(0) complex subunit C3, mitochondrial	0.608176
Q15046	KARS	Lysine--tRNA ligase	0.608522
P00338	LDHA	L-lactate dehydrogenase A chain	0.615549
P36873	PPP1CC	Serine/threonine-protein phosphatase PP1-gamma catalytic subunit	0.620001
Q15366	PCBP2	Poly(rC)-binding protein 2	0.622055
P49770	EIF2B2	Translation initiation factor eIF-2B subunit beta	0.625432
P04844	RPN2	Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 2	0.637911
P0DP23	CALM1	Calmodulin-1	0.648649
P14406	COX7A2	Cytochrome c oxidase subunit 7A2, mitochondrial	0.656231
P14854	COX6B1	Cytochrome c oxidase subunit 6B1	0.657215
P62875	POLR2L	DNA-directed RNA polymerases I, II, and III subunit RPABC5	0.675001
P0CG48	UBC	Polyubiquitin-C	0.677235
P10644	PRKAR1A	cAMP-dependent protein kinase type I-alpha regulatory subunit	0.680635
O15144	ARPC2	Actin-related protein 2/3 complex subunit 2	0.693822
P17844	DDX5	Probable ATP-dependent RNA helicase DDX5	0.73098
O95467	GNAS	Neuroendocrine secretory protein 55	0.731569
P63092	GNAS	Guanine nucleotide-binding protein G(s) subunit alpha isoforms short	0.731569
P84996	GNAS	Protein ALEX	0.731569

Inference of pre-processing transformations

Text mining

The metadata description of the MA samples was processed using a text mining pipeline to build datasets that would allow the training of $ML_{processing}$ models to predict pre-processing procedures applied to the samples. Given the complexity of some of the sentences, descriptions comprising more than two phrases were excluded. With the remaining examples, we queried the samples for “hot words”: “calibrat”, which is a simplified word that allows the detection of various verbal forms of calibration; “intensity” ; “log” ; “log 10” and “log10” ; “log 2” and “log2”, “mas”, “normali”, a simplified word that allows the detection of American and British

English, as well as various verbal forms normalization, “quantile”, “ratio”, “raw” ; and “rma”. The final/resulting dataset will include samples which included one or more “hot words”, as portrayed in **Table 4**. A random selection of 5000 samples following the rules in **Table 4** was recorded and joined with the expression-related features previously described to compose the datasets used for predicting pre-processing of samples (sections below). 20% of the samples were reserved for benchmarking of the trained models.

Table 4. **Composition of datasets after text mining procedures.** The “+” on the right illustrated the “applicable” class, and the “-” the “non-applicable” class. The binary matrix indicates if the samples were included depending on the “hot word”. Cells with “1” indicate that samples with the word were included, cells with “0” indicate that the samples were excluded, and cells with “-” indicate that the “hot word” was ignored. * The term “Normalized” is an umbrella term that additionally includes the words “normali”, “mas”, “quantile”, and “rma”.

Dataset	**Log2**	**Log10**	**Log**	**Ratio**	**Raw**	**Normalized***
Log2+	1	0	-	-	-	-
Log2-	0	1	-	-	-	-
Log10+	0	1	-	-	-	-
Log10-	1	0	-	-	-	-
Ratio+	-	-	-	1	0	-
Ratio-	-	-	-	0	1	-
Log+	-	-	1	-	0	-
Log-	-	-	0	-	1	0
Normalized+	-	-	-	-	0	1
Normalized-	-	-	-	-	1	0

ML models for pre-processing inference

Model Architecture

The datasets described in the methods’ sections above were used to train $ML_{processing}$ classifiers to predict what methods had been applied as data pre-processing when these details were unknown. We used a set of different ML algorithms and selected the best-performing ones for application in the following steps: Decision tree[85], Elastic net[86], Elastic net with threshold optimization[87], Generalized linear model[88], Linear classifier[89], Multi-layer perceptron[90], Naïve Bayes Classifier[91], Quadratic classifier[89], and Support Vector Machines[92]. Additionally, in each of the models, we also carried out a feature selection process based upon a diversified set of methods described in **Supplementary Table 8**. Most models were included in MATLAB (version 9.2)[93] and its “Statistics and Machine Learning Toolbox”[94]. Models that were not available (Linear and Quadratic classifiers, and threshold

optimization) were deployed using custom MATLAB code. The hyperparameters of the applied models are described in **Table 5**. The threshold optimization step in Elastic net with threshold optimization models was done by calculation of the receiver operator curve (ROC). The final models used for the prediction of MA samples' pre-processing consists of soft-voting ensemble models comprising the top-three best-performing models in each case/dataset.

Table 5. ML models hyperparameters. Parameters not mentioned in this table were left as default.

Model	**Hyperparameters**	***Model***	**Hyperparameters**
Decision tree	Prune ratio: 0; Minimum samples per leaf: 1;	*Multi-layer perceptron*	No hidden layers: 1; No hidden nodes: 2; No epochs: 200; Activation function: Tangent sigmoid;
Elastic net	Alpha = 0.01	*Naïve Bayes Classifier*	-
Generalized linear model	Distribution: binomial; Link: Normal inverse;	*Support Vector Machines*	-

Validation of pre-processing inference $ML_{processing}$ models was performed by splitting the dataset (defined in sections above and according to **Table 4**) into training and test sets~~.~~ 20% of samples were reserved for benchmarking, and the remaining 80% were used for training. Training was done in a 25-k cross-validation strategy. The models showing the best training metrics were kept and used in the remaining steps.

Prediction Confidence

To improve the $ML_{processing}$ models' confidence, the raw score (model output before the output activation functions) from each model in the various ensembles was recorded. Scores were split into 5%-quantile intervals and the mean accuracy was calculated for each of them. This value was used to check the models' certainty level depending on the raw score value. Low-confidence regions were disregarded and only predictions falling into intervals with more than 95% mean accuracy were considered (figure S1).

Uniformization of processing techniques

Samples were classified by $ML_{processing}$ models. In cases where pre-processing was unknown/unclear, were transformed according to classification results. The transformations applied are described in equations 9 and 10.

Where log2 is detected,

$$x' = 2^x \quad \text{(Eq. 9)}$$

And, where log10 is detected,

$$x' = 10^x \quad \text{(Eq. 10)}$$

Cross-sample normalization method

After undoing logarithm transformations, we applied a normalization technique to uniformly transform all samples in the repository. Given that different samples may have been acquired in different settings (e.g., different MA scanners, processing software), and that we could not resort to control measurements in the samples (e.g., control probes), we decided to apply a normalization strategy that would level the various samples using a biologically intrinsic value in the samples. We have used the intensity values of HK genes to infer a sample-specific normalization curve that would be capable of normalizing the expression of the remaining genes. The intensity values from the least variable 40 HK genes (in Table 3) were statistically described across all samples in the repository. Given that none of the genes show a normal distribution of intensity, we have used a cross-GEO background dataset containing the median expression values of these 40 genes (described in **Supplementary Table 9**). For each sample, we built a dataset in which the median HK expression values were mapped to the expression value in the given sample. This sample-specific dataset was used to train two types of regressors – (1) spline[95] and (2) polynomial regressors[96] – that were applied to the remaining genes

in the sample dataset. As results from both regressions were equivalent, future steps were performed using the polynomial regression for its stronger theoretical support.

Benchmark of MACAROON against other integration methods

We have tested MACAROON against a YuGene[49] – the only other MA processing method capable of handling pre-processed data, as means of evaluating its performance in comparison to an original MA dataset (GEO series GSE102124[50] – corresponding to the "Gene expression profiling of treated and untreated primary prostate cancer"). The three datasets were submitted to a DE analysis where expression values between treated and untreated samples were compared. P-values were calculated using a student's t-test.

Comparative case study

To evaluate the performance of MACAROON, we collected 200 samples of Prostate Cancer from 63 GEO series (see **Supplementary Table 3** for further detail). To account for possible errors in the ML-based stages of MACAROON, maximum expression values in each sample were assessed, and samples corresponding to outlier maximum values were excluded (applying the interquartile range criterion[97]). DE analysis resulted from the comparison between intervention vs. control samples in the corresponding GEO series. DEGs were selected using a threshold |logFC|>=0.5 and FDR <0.05, using a Student's T-test. DEGs were analysed for Gene Ontology BP[52] enrichment using PANTHER[53].

Metadata mining and RWD acquisition

In order to accomplish the second objective of this work – to build an RWD repository from GEO datasets – we mined metadata information provided alongside GEO samples. Medical conditions and demographical information were extracted from each entry. Additionally, knowing the significant differences existing between gene expression across tissues, samples'

tissue information was also retrieved. Information was mined from the text descriptions of GEO samples.

Demographical features

To assign the samples' demographic characteristics, we again mined the provided metadata text for information on age, biological sex, ethnicity, BMI and height. Age was extracted mainly from the subfields "age" and "birth". The obtained values were then standardized to comprise the age of the subject in years. Biological sex was mined from the subfield "gender". Different available classes (given that sex is described using different notations) were standardized to "female" and "male". Ethnicity was obtained from the fields "ethnicity", "race" and "population". The identified vocabulary was standardized to "african", "asiatic", "caucasian", "hispanic" and "indian". Information on BMI was mined from the subfield "BMI", with no further transformations. Height values were extracted from the subfield "height". Units were mined from text and converted to centimetres (when values were presented in other units).

Medical conditions

Samples were tagged according to the medical condition provided in GEO series and the samples' titles and descriptions. Medical conditions, corresponding to BED terms, were assigned to each sample by using a manually curated synonym dictionary where metadata sentences were mined for a series of hot words. Additionally, healthy and control samples were explicitly mined to account for different experimental designs in the GEO series. The different BED terms were mapped to MeSH terms[77] using MetaMap[98] – a tool to map biomedical terms referred to in raw text to controlled terms in the Unified Medical Language System (UMLS) and corresponding condition ids (CUI). This allowed the hierarchical grouping of the different medical conditions for grouped analysis.

Information on source tissue

To account for expression differences, information regarding the samples' body tissues was also extracted from the metadata file. To reduce redundancy, and to allow for hierarchical classification of body tissues, terms were mapped to the Brenda Tissue Ontology[99].

ML models to predict BED medical conditions

For each BED medical condition mined from the GEO samples, we trained a set of $ML_{clinical}$ models aimed at deciphering further medical conditions in the samples. Proteins associated with each medical condition in the BED were selected as features. Models were trained in a grid-search fashion, including feature selection (following the strategy applied to $ML_{processing}$). A maximum of 2,000 samples were used for each class, with the positive class containing GEO samples from which the specific BED medical condition was mined. On the other side, the negative class corresponded to a random set of samples where that medical condition was not mentioned. To account for variability in the number of samples, models were selected based on their cross-validation (k=10) BACC, where only the models with balanced accuracy above 80% (p-value<0.01) were kept. To account for missing values, for each medical condition, models had a different number of features (ranging from 2 to 5).

Abbreviation

ACC	Accuracy
AI	Artificial Intelligence
BED	Biological Effectors Database
BMI	Body Mass Index
BP	Biological Process
CUI	Condition ID

1	CV	Cross-validation
2	DE	Differential Expression
3	DEG	Differentially Expressed Genes
4	EMA	European Medicines Agency
5	FC	Fold-change
6	FDA	Food and Drug Administration
7	FN	False Negatives
8	FP	False Positives
9	GEO	Gene Expression Omnibus
10	HK	House-Keeping
11	HT	High-throughput
12	MA	Microarrays
13	MeSH	Medical Subject Headings
14	ML	Machine Learning
15	NLP	Natural Language Processing
16	PRE	Precision
17	RCT	Randomized Clinical Trial
18	RNA-Seq	RNA Sequencing
19	ROC	Receiver-operator Curve
20	RWD	Real-World Data
21	SNS	Sensitivity

SPC	Specificity
TN	True Negatives
TP	True Positives
UMLS	Unified Medical Language System

DECLARATIONS

Ethics approval and consent to participate

The patient data included in this work have been obtained from GEO database, where cell lines and other sources of information not adequate for our proposes were excluded. GEO is a public repository that contains molecular data that could be considered sensible from certain ethical perspectives given that these data describe the molecular status of patients. However, this source of data does not contain personal details or any other way to allow the patient identification and neither other clinical detail of them. Despite this fact and to reinforce the patients' privacy, all patients included in this study were submitted to a pseudonymisation protocol to avoid any possible link between our patient repository and the original GEO identification. This data treatment has been done according to the EU regulation on personal data and its management (specifically, GDPR – EU regulation 2016/679). Additionally, we also choose not to provide our patient repository to avoid any possible misuse by third people or entities about the clinical information estimated during our analyses and to ensure the application of the "appropriate safeguards" according to this regulation. All measures taken in this study are driven to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject. The consent of the participants has been treated according to the exposed regulations and data management and considering that this study is a non-interventional work

with scientific and public interest, based on historical research and with statistical purposes that show aggregated information to extract the biological and clinical conclusions.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are publicly available in the GEO repository, the whole database is available at https://ftp.ncbi.nlm.nih.gov/geo/. Series applied in benchmarking and case studies are referenced in the methods and references sections; and in the supplementary material.

Competing interests

The corresponding author declares that there is no conflict of interest with the authors.

Funding

PMF, JMGI, GJ and JF are employees of Anaxomics Biotech S.L. JMM is an employee and co-founder of Anaxomics Biotech S.L. BO is a scientific advisor for Anaxomics Biotech but does not perceive any funding support for these services. PMF and JMGI received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860303 (proEVLifeCycle) and 859962 (SECRET), respectively.

Authors' contributions

PMF, methodology, software, validation, formal analysis, investigation, resources, Writing – original draft, Writing – review & editing, visualization. JMGI, methodology, data curation, formal analysis, Writing – review & editing. GJ, conceptualization, methodology, Validation,

formal analysis, data curation, Writing – review & editing. BO, Supervision, Writing – review & editing. JF, Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Validation, Writing – review & editing. JMM, Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Acknowledgements

Not applicable.

REFERENCES

1. Jayatunga MKP, Xie W, Ruder L, Schulze U, Meier C (2022) AI in small-molecule drug discovery: a coming wave? Nat Rev Drug Discov 21:175–176

2. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK (2021) Artificial intelligence in drug discovery and development. Drug Discov Today 26:80–93

3. Koscielny G, An P, Carvalho-Silva D, et al (2017) Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res 45:D985–D994

4. Jiménez-Luna J, Grisoni F, Schneider G (2020) Drug discovery with explainable artificial intelligence. Nature Machine Intelligence 2020 2:10 2:573–584

5. Rodriguez S, Hug C, Todorov P, et al (2021) Machine learning identifies candidates for drug repurposing in Alzheimer's disease. Nature Communications 2021 12:1 12:1–13

6. Rashid MBMA, Toh TB, Hooi L, et al (2018) Optimizing drug combinations against multiple myeloma using a quadratic phenotypic optimization platform (QPOP). Sci Transl Med. https://doi.org/10.1126/SCITRANSLMED.AAN0941/SUPPL_FILE/AAN0941_TABLE_S19.ZIP

7. Kolla L, Gruber FK, Khalid O, Hill C, Parikh RB (2021) The case for AI-driven cancer clinical trials – The efficacy arm in silico. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer 1876:188572

8. Collins GS, Dhiman P, Andaur Navarro CL, et al (2021) Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 11:e048008

9. Baumfeld Andre E, Reynolds R, Caubel P, Azoulay L, Dreyer NA (2020) Trial designs using real-world data: The changing landscape of the regulatory approval process. Pharmacoepidemiol Drug Saf 29:1201–1212

10. (2018) Framework for FDA's Real-World Evidence Program.

11. Cave A, Kurz X, Arlett P (2019) Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe. Clin Pharmacol Ther 106:36–39

12. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices | FDA. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices. Accessed 25 Mar 2022

13. Zura R, Irwin DE, Mack CD, Aldridge ML, Mackowiak JI (2021) Real-World Evidence: A Primer. J Orthop Trauma. https://doi.org/10.1097/BOT.0000000000002037

14. Crick F (1970) Central Dogma of Molecular Biology. Nature 1970 227:5258 227:561–563

15. Shapiro JA (2009) Revisiting the central dogma in the 21st century. Ann N Y Acad Sci 1178:6–28

16. Melé M, Ferreira PG, Reverter F, et al (2015) The human transcriptome across tissues and individuals. Science (1979) 348:660–665

17. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Hieter P, Vogelstein B, Kinzler KW (1997) Characterization of the Yeast Transcriptome. Cell 88:243–251

18. Bainbridge MN, Warren RL, Hirst M, et al (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics 7:1–11

19. Fu X, Fu N, Guo S, et al (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics 10:1–9

20. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X (2014) Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLoS One 9:e78644

21. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57

22. Minnier J, Pennock ND, Guo Q, Schedin P, Harrington CA (2018) RNA-seq and expression arrays: Selection guidelines for genome-wide expression profiling. Methods in Molecular Biology 1783:7–33

23. Thompson JA, Tan J, Greene CS (2016) Cross-platform normalization of microarray and RNA-seq data for machine learning applications. PeerJ 2016:e1621

24. Belle A, Thiagarajan R, Soroushmehr SMR, Navidi F, Beard DA, Najarian K (2015) Big data analytics in healthcare. Biomed Res Int. https://doi.org/10.1155/2015/370194

25. Roski J, Bo-Linn GW, Andrews TA (2014) Creating value in health care through big data: Opportunities and policy implications. Health Aff 33:1115–1122

26. Chen L, Sun F, Yang X, Jin Y, Shi M, Wang L, Shi Y, Zhan C, Wang Q (2017) Correlation between RNA-Seq and microarrays results using TCGA data. Gene 628:200–204

27. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, Nielsen J (2012) A comprehensive comparison of RNA-Seq-based transcriptome analysis from

reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res 40:10084–10097

28. Park SH, Choi J, Byeon JS (2021) Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence. Korean J Radiol 22:442–453

29. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE (2015) Big Data: Astronomical or Genomical? PLoS Biol 13:e1002195

30. Barrett T, Wilhite SE, Ledoux P, et al (2013) NCBI GEO: Archive for functional genomics data sets - Update. Nucleic Acids Res. https://doi.org/10.1093/nar/gks1193

31. Cheng L, Lo LY, Tang NLS, Wang D, Leung KS (2016) CrossNorm: a novel normalization strategy for microarray data in cancers. Scientific Reports 2016 6:1 6:1–11

32. Junet V, Farrés J, Mas JM, Daura X (2021) CuBlock: a cross-platform normalization method for gene-expression microarrays. Bioinformatics 37:2365–2373

33. Carey M, Ramírez JC, Wu S, Wu H (2018) A big data pipeline: Identifying dynamic gene regulatory networks from time-course Gene Expression Omnibus data with applications to influenza infection. Stat Methods Med Res 27:1930–1955

34. Koeppen K, Stanton BA, Hampton TH (2017) ScanGEO: parallel mining of high-throughput gene expression data. Bioinformatics 33:3500–3501

35. Barrett T, Edgar R (2006) [19] Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis. Methods Enzymol 411:352–369

36. Chen G, Ramírez JC, Deng N, Qiu X, Wu C, Zheng WJ, Wu H (2019) Restructured GEO: restructuring Gene Expression Omnibus metadata for genome dynamics analysis. Database 2019:145

37. Khomtchouk BB, Dyomkin V, Vand KA, Assimes T, Gozani O (2018) Biochat: a database for natural language processing of Gene Expression Omnibus data. bioRxiv 480020

38. Hadley D, Pan J, El-Sayed O, et al (2017) Precision annotation of digital samples in NCBI's gene expression omnibus. Scientific Data 2017 4:1 4:1–11

39. Brazma A, Parkinson H, Sarkans U, et al (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31:68–71

40. Fukuda A, Kodama Y, Mashima J, Fujisawa T, Ogasawara O (2021) DDBJ update: streamlining submission and access of human data. Nucleic Acids Res 49:D71–D75

41. Lappalainen I, Almeida-King J, Kumanduri V, et al (2015) The European Genome-phenome Archive of human data consented for biomedical research. Nature Genetics 2015 47:7 47:692–695

42. Arisi I, Dronofrio M, Brandi R, et al (2019) Mining clinical and laboratory data of neurodegenerative diseases by Machine Learning: Transcriptomic biomarkers. Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018 2735–2737

43. Kato S, Kim KH, Lim HJ, et al (2020) Real-world data from a molecular tumor board demonstrates improved outcomes with a precision N-of-One strategy. Nature Communications 2020 11:1 11:1–9

44. Fernandes LE, Epstein CG, Bobe AM, et al (2021) Real-world Evidence of Diagnostic Testing and Treatment Patterns in US Patients With Breast Cancer With Implications

for Treatment Biomarkers From RNA Sequencing Data. Clin Breast Cancer 21:e340–e361

45. Wallach JD, Zhang AD, Skydel JJ, Bartlett VL, Dhruva SS, Shah ND, Ross JS (2021) Feasibility of Using Real-world Data to Emulate Postapproval Confirmatory Clinical Trials of Therapeutic Agents Granted US Food and Drug Administration Accelerated Approval. JAMA Netw Open 4:e2133667–e2133667

46. Refaeilzadeh P, Tang L, Liu H (2009) Cross-Validation. Encyclopedia of Database Systems 532–538

47. Zahurak M, Parmigiani G, Yu W, Scharpf RB, Berman D, Schaeffer E, Shabbeer S, Cope L (2007) Pre-processing Agilent microarray data. BMC Bioinformatics 8:1–13

48. Vafaee F, Dashti H, Alinejad-Rokny H (2019) Transcriptomic Data Normalization. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (eds) Encyclopedia of Bioinformatics and Computational Biology. Academic Press, pp 364–371

49. Lê Cao KA, Rohart F, McHugh L, Korn O, Wells CA (2014) YuGene: A simple approach to scale gene expression data derived from different platforms for integrated analyses. Genomics 103:239–251

50. Sowalsky AG, Ye H, Bhasin M, et al (2018) Neoadjuvant-Intensive Androgen Deprivation Therapy Selects for Prostate Tumor Foci with Diverse Subclonal Oncogenic Alterations. Cancer Res 78:4716–4730

51. Edwards D (2003) Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics 19:825–833

52. Carbon S, Douglass E, Good BM, et al (2021) The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 49:D325–D334

53. Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, Thomas PD (2021) PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res 49:D394–D403

54. Yang J, Chen L, Kong X, Huang T, Cai YD (2014) Analysis of Tumor Suppressor Genes Based on Gene Ontology and the KEGG Pathway. PLoS One 9:e107202

55. Fan S, Liang Z, Gao Z, Pan Z, Han S, Liu X, Zhao C, Yang W, Pan Z, Feng W (2018) Identification of the key genes and pathways in prostate cancer. Oncol Lett 16:6663–6669

56. Sun J, Li S, Wang F, Fan C, Wang J (2019) Identification of key pathways and genes in pten mutation prostate cancer by bioinformatics analysis. BMC Med Genet 20:1–9

57. Verma S, Shukla S, Pandey M, MacLennan GT, Gupta S (2019) Differentially expressed genes and molecular pathways in an autochthonous mouse prostate cancer model. Front Genet 10:235

58. Montes J, McDermott MP, Mirek E, et al (2018) Ambulatory function in spinal muscular atrophy: Age-related patterns of progression. PLoS One 13:e0199657

59. Przemska-Kosicka A, Childs CE, Maidens C, Dong H, Todd S, Gosney MA, Tuohy KM, Yaqoob P (2018) Age-related changes in the natural killer cell response to seasonal influenza vaccination are not influenced by a synbiotic: A randomised controlled trial. Front Immunol 8:591

60. Basoglu OK, Tasbakan MS (2018) Gender differences in clinical and polysomnographic features of obstructive sleep apnea: a clinical study of 2827 patients. Sleep and Breathing 22:241–249

61. Gabler NB, French B, Strom BL, Liu Z, Palevsky HI, Taichman DB, Kawut SM, Halpern SD (2012) Race and Sex Differences in Response to Endothelin Receptor Antagonists for Pulmonary Arterial Hypertension. Chest 141:20–26

62. Morris JC, Schindler SE, McCue LM, et al (2019) Assessment of Racial Disparities in Biomarkers for Alzheimer Disease. JAMA Neurol 76:264–273

63. Kopel J, Perisetti A, Roghani A, Aziz M, Gajendran M, Goyal H (2020) Racial and Gender-Based Differences in COVID-19. Front Public Health 8:418

64. Hebbes CP, Thompson JP (2018) Pharmacokinetics of anaesthetic drugs at extremes of body weight. BJA Educ 18:364–370

65. Morrish GA, Pai MP, Green B (2011) The effects of obesity on drug pharmacokinetics in humans. http://dx.doi.org/101517/174252552011570331 7:697–706

66. Smit C, de Hoogd S, Brüggemann RJM, Knibbe CAJ (2018) Obesity and drug pharmacology: a review of the influence of obesity on pharmacokinetic and pharmacodynamic parameters. https://doi.org/101080/1742525520181440287 14:275–285

67. Tittl L, Endig S, Marten S, Reitter A, Beyer-Westendorf I, Beyer-Westendorf J (2018) Impact of BMI on clinical outcomes of NOAC therapy in daily care - Results of the prospective Dresden NOAC Registry (NCT01588119). Int J Cardiol 262:85–91

68. Donini LM, Pinto A, Giusti AM, Lenzi A, Poggiogalle E (2020) Obesity or BMI Paradox? Beneath the Tip of the Iceberg. Front Nutr 7:53

69. Hadley D, Pan J, El-Sayed O, et al (2017) Precision annotation of digital samples in NCBI's gene expression omnibus. Scientific Data 2017 4:1 4:1–11

70. Shah N, Guo Y, Wendelsdorf K v., Lu Y, Sparks R, Tsang JS (2016) A crowdsourcing approach for reusing and meta-analyzing gene expression data. Nature Biotechnology 2016 34:8 34:803–806

71. Bernstein MN, Doan A, Dewey CN (2017) MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. Bioinformatics 33:2914–2923

72. Giles CB, Brown CA, Ripperger M, Dennis Z, Roopnarinesingh X, Porter H, Perz A, Wren JD (2017) ALE: Automated label extraction from GEO metadata. BMC Bioinformatics 18:7–16

73. Wang Z, Lachmann A, Ma'ayan A (2019) Mining data and metadata from the gene expression omnibus. Biophys Rev 11:103–110

74. Jorba G, Aguirre-Plans J, Junet V, Segú-Vergés C, Ruiz JL, Pujol A, Fernández-Fuentes N, Mas JM, Oliva B (2020) In-silico simulated prototype-patients using TPMS technology to study a potential adverse effect of sacubitril and valsartan. PLoS One 15:e0228926

75. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 48:D845–D855

76. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M (2021) KEGG: integrating viruses and cellular organisms. Nucleic Acids Res 49:D545–D551

77. National Center for Biotechnology Information, U.S. National Library of Medicine MeSH (Medical Subject Headings). https://www.ncbi.nlm.nih.gov/mesh/. Accessed 30 Nov 2021

78. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210

79. Richardson L (2007) Beautiful soup documentation.

80. Butte AJ, Dzau VJ, Glueck SB (2001) Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues." https://doi.org/101152/physiolgenomics20017295 7:95–96

81. Zhu J, He F, Hu S, Yu J (2008) On the nature of human housekeeping genes. Trends in Genetics 24:481–484

82. Greer S, Honeywell R, Geletu M, Arulanandam R, Raptis L (2010) Housekeeping genes; expression levels may change with density of cultured cells. J Immunol Methods 355:76–79

83. Eisenberg E, Levanon EY (2013) Human housekeeping genes, revisited. Trends in Genetics 29:569–574

84. Zhang Y, Li D, Sun B (2015) Do Housekeeping Genes Exist? PLoS One 10:e0123691

85. Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) Classification and regression trees. Classification and Regression Trees 1–358

86. Zou H, Hastie T (2005) Regularization and Variable Selection via the Elastic Net. J R Stat Soc Series B Stat Methodol 67:301–320

87. Boughorbel S, Jarray F, El-Anbari M (2017) Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS One 12:e0177678

88. Madsen H, Thyregod P (2010) Introduction to general and generalized linear models. Introduction to General and Generalized Linear Models. https://doi.org/10.1201/9781439891148/INTRODUCTION-GENERAL-GENERALIZED-LINEAR-MODELS-HENRIK-MADSEN-POUL-THYREGOD

89. Fukunaga K (1990) Introduction to Statistical Pattern Recognition. Introduction to Statistical Pattern Recognition. https://doi.org/10.1016/C2009-0-27872-X

90. Haykin S (1998) Neural Networks: A Comprehensive Foundation, 2nd ed. Prentice Hall PTR, USA

91. Russell SJ, Norvig P (2010) Artificial Intelligence: A Modern Approach, 3rd ed. Pearson Education, Inc., New Jersey

92. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G (2008) Support Vector Machines and Kernels for Computational Biology. PLoS Comput Biol 4:e1000173

93. The Mathworks Inc (2017) MATLAB: r2017b.

94. Mathworks Statistics and Machine Learning Toolbox Documentation.

95. (2017) MATLAB Smoothing Splines.

96. (2017) MATLAB Polynomial Models.

97. Dekking FM, Kraaikamp C, Lopuhaä HP, Meester LE (2005) Exploratory data analysis: numerical summaries. 231–243

98. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32:D267–D270

99. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, Schomburg D (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Res 39:D507–D513

The usage of transcriptomics datasets as sources of Real-World Data for clinical trialling

Figures list

Pedro Matos-Filipe, Juan Manuel García-Illarramendi, Guillem Jorba, Baldo Oliva, Judith Farrés, José Manuel Mas

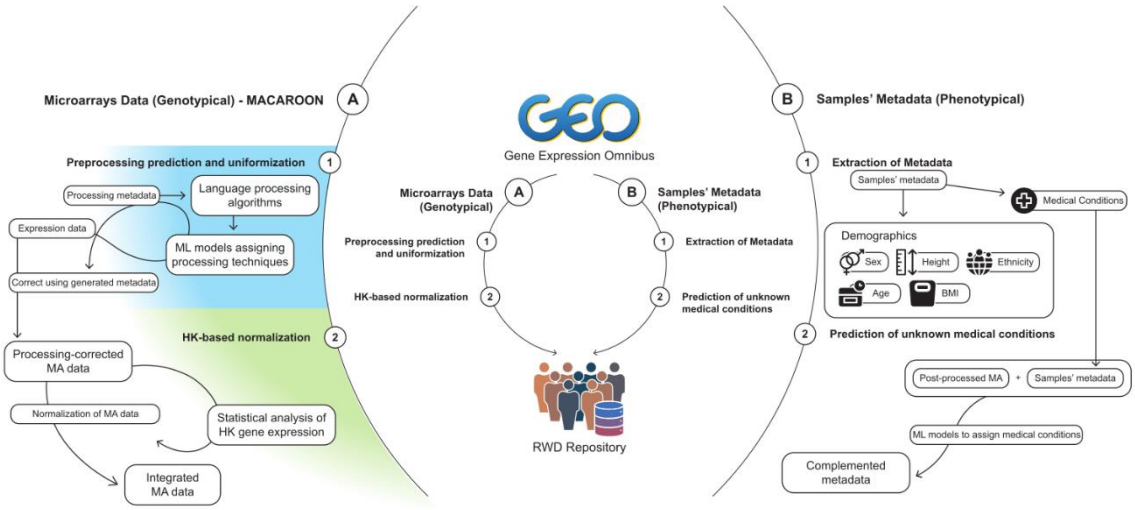

Figure 1. The overall pipeline of the methods applied in this work. A. Collection and processing of MA data. B. Collection and processing of RWD from GEO.

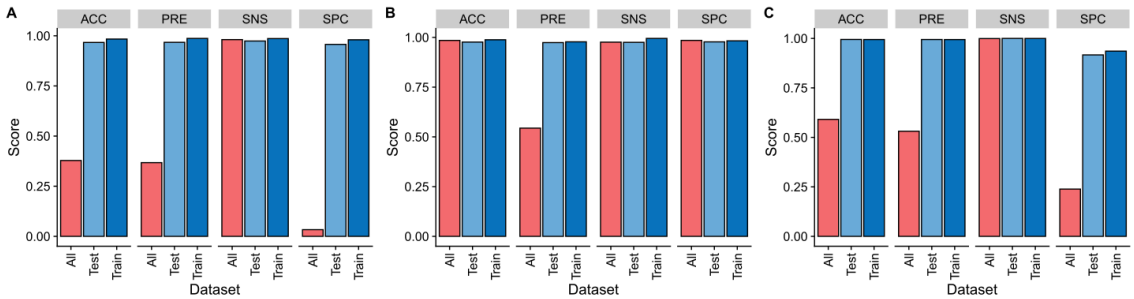

Figure 2. Global test metrics of the various ensemble models in $ML_{processing}$. (A) Log2, (B) Log10, and (C) Ratio

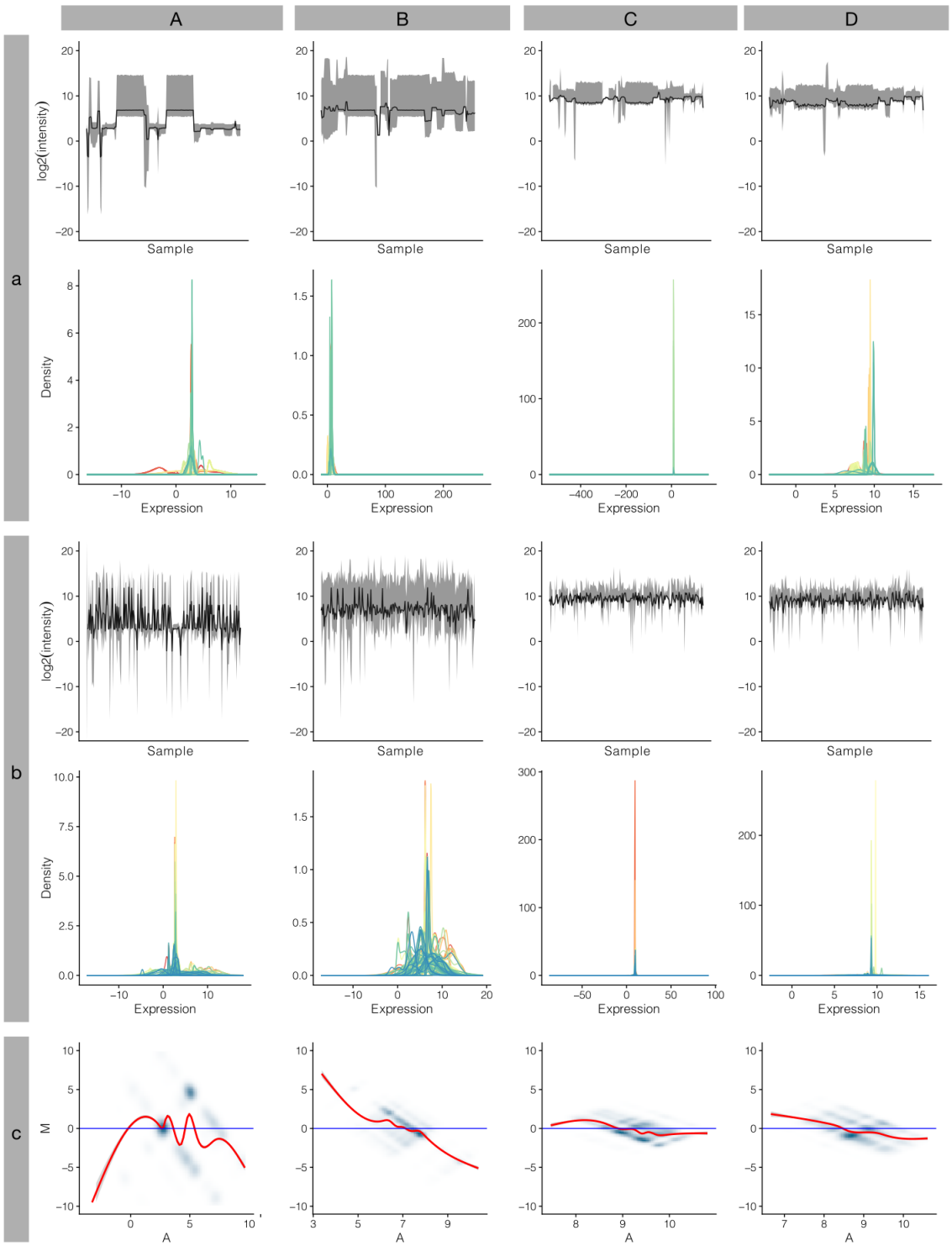

Figure 3. Log-expression data distribution across the various stages of MACAROON. A, B, C and D refer to the datasets produced along the process: A. Raw unprocessed data; B. After reverting preprocessing; C. After spline regression; and D. After polynomial regression. (a), (b), and (c) concern the samples displayed in the figure: (a). healthy control samples; (b). prostate cancer samples; and (c). MA-plot (control vs. prostate cancer) – as described in Eq. 5 and 6.

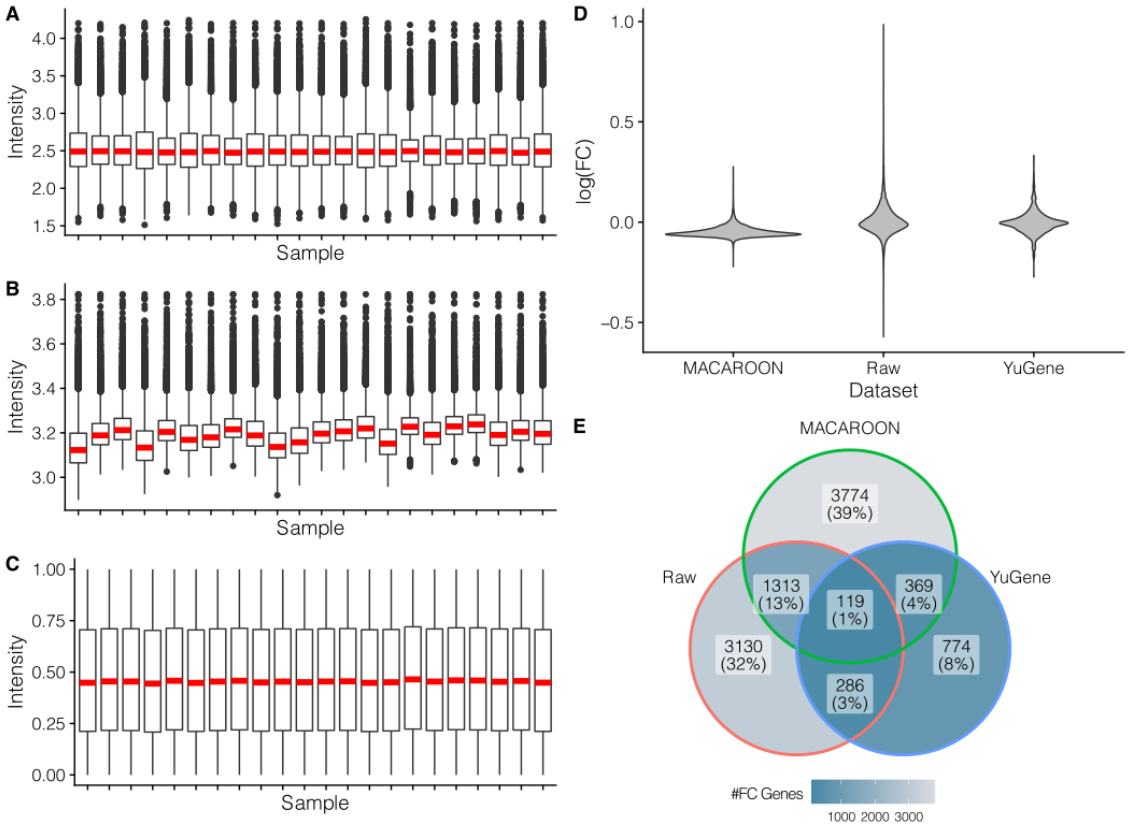

Figure 4. Direct comparison between methods. In the left column, we show the intensity per sample. A corresponds to the log2 raw data; B, using MACAROON; and C, using YuGene[49]. In the right column, we display a comparison of the results of a Differential Expression analysis. In D, we show a dispersion of the log Fold Change values; in E a Venn diagram portraying common differentially expressed genes depending on the dataset used.

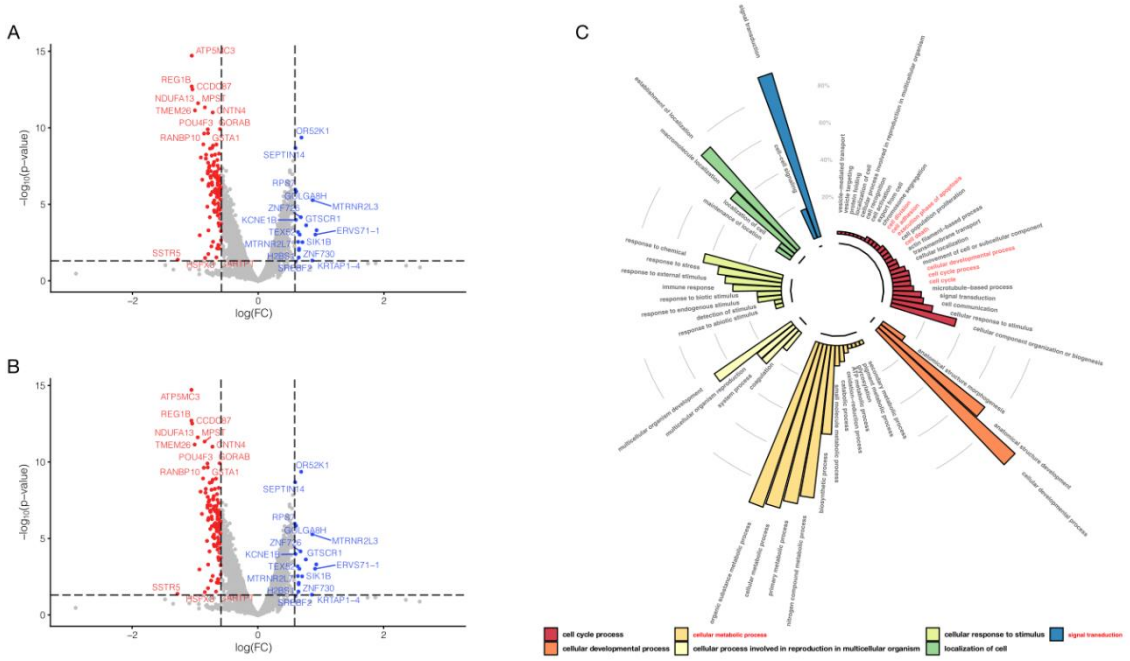

Figure 5. DE analysis results using spline and polynomial regressions. A, B. Volcano plots of the DE using both regression techniques. C. Enrichment analysis results using the polynomial dataset (where BPs found in other works in literature are highlighted in red). In A and B, the dots in red are under-expressed genes and the dots in blue are over-expressed genes (prostate cancer vs. control). In C, colours represent the first rank of GO BPs, while each column refers to the second rank.

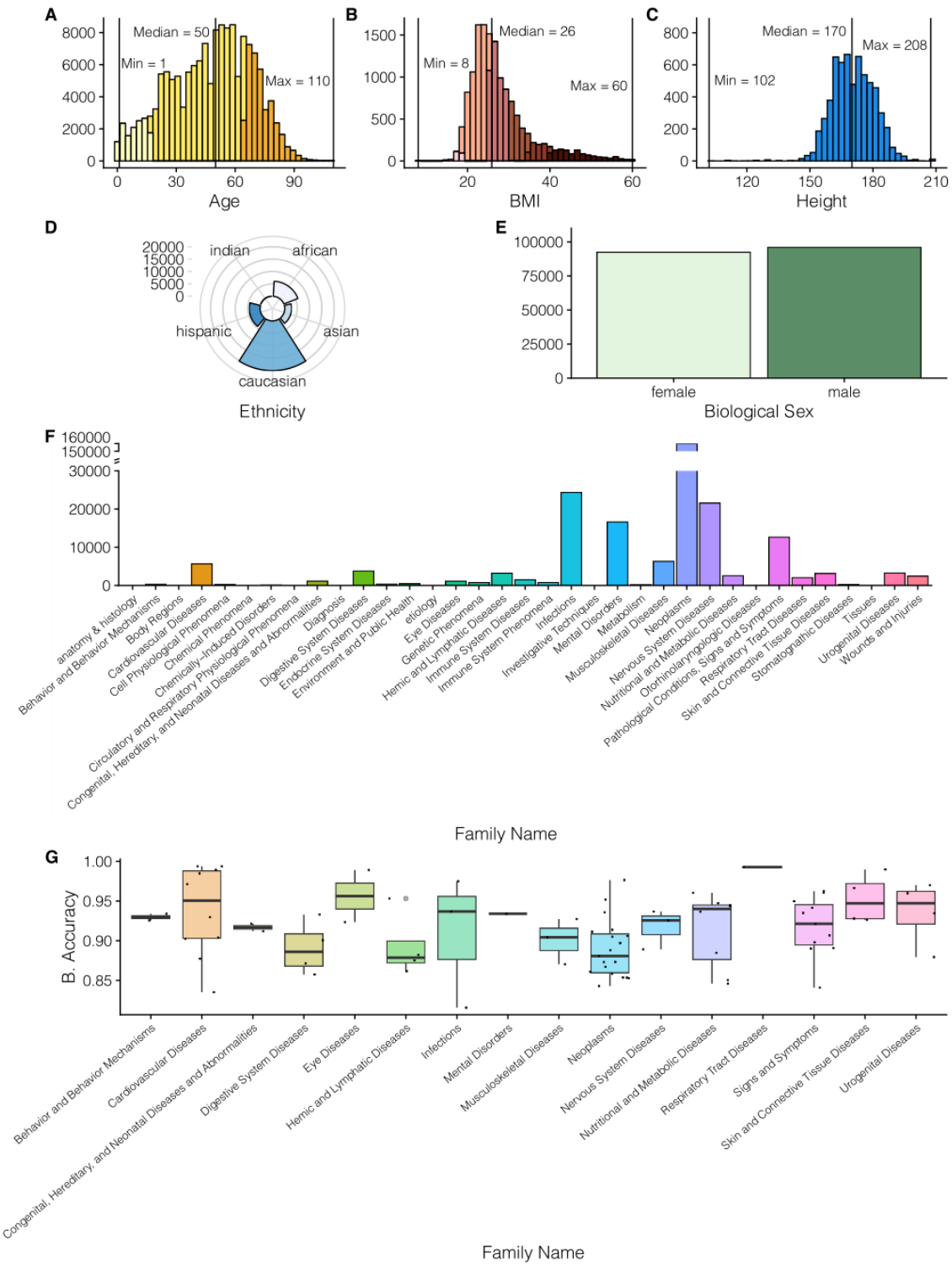

Figure 6. Demographic description of MA samples in GEO. A, Age distribution of samples (n=152,237). Colours represent various age groupings: light yellow – childhood; yellow – adulthood; orange – elderhood. B, BMI distribution of samples (n=14,243). Different shades of red represent different gradings of BMI: very light – underweight; light – normal; red – pre-obese; dark– obese class I; very dark – obese class II. C, Height distribution of samples (n=6,200). D, Ethnicity background ratios (n=33,465). E, Biological sex ratio of samples (n=188,618). F, MeSH terms of disease conditions (top leaf in the MeSH ontology tree) found in GEO samples. G, balanced accuracy of the $ML_{clinical}$ models aimed at predicting samples' medical conditions. Results are grouped by the MeSH family. Each black dot corresponds to the average balanced accuracy of a medical condition in that family.