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*]

[1*]Department, Anaxomics Biotech SL, Carrer Diputaciò 237, Barcelona, 08003, Catalonia, Spain.
[2]Structural Bioinformatics Group, Research Programme on Biomedical Informatics, Department of Experimental and Health Science, Universitat Pompeu Fabra, Doctor Aiguader 80, Barcelona, 10587, Catalonia, Spain.
[3]Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Carrer de la Vinya, Cerdanyola del Vallès, 08193, Catalonia, Spain.

*Corresponding author(s). E-mail(s): baldo.oliva@upf.edu; xouse@anaxomics.com;

Abstract

Background: Randomised Clinical Trials (RCT) reflect results within their specific controlled settings, necessitating further studies to understand outcomes across all possible scenarios. The usage of Real-World Data (RWD) has been recently considered to be a viable alternative to overcome these issues and complement clinical conclusions. Molecular profiles of patients captured by high-throughput measures reflect their medical conditions. When this information is linked to clinical and demographical information, nuances in transcriptomics data can uncover subtle variations in disease pathways among distinct patient groups. This work focuses on the construction of a patient repository database with molecular and 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. Our post-processing technique, which we have named MicroArray Cross-plAtfoRm

pOst-prOcessiNg (MACAROON), aims to standardise 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 45% improvement compared to other methods available. Furthermore, RWD was mined from GEO samples' metadata and a clinical and demographical characterisation of the database was obtained. 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 a repository, which includes molecular, clinical and demographical RWD by integrating multiple public datasets. The exploration of these data facilitates the discovery of clinical outcomes and molecular pathways specific to predetermined patient populations.

Keywords: RWD; Data Integration, Big Data, Microarrays, Post-processing, Normalisation

1 Introduction

Developments in the fields of Artificial Intelligence (AI) and Machine Learning (ML) (and their association with the pharmaceutical industry)have helped to boost the time-line of production of pharmaceutical products, leading to a reduction in cost (Jayatunga et al., 2022). Techniques in the field are already being streamlined into production under various stages of the life cycle of pharmaceutical products (Paul et al., 2021), namely, drug discovery (Koscielny et al., 2017; Jiménez-Luna et al., 2020; Rodriguez et al., 2021), pharmaceutical development (Rashid et al., 2018), and clinical trial development and monitoring (Kolla et al., 2021; Collins et al., 2021). Nonetheless, regulatory agencies still mostly admit evidence acquired through Randomised Clinical Trials (RCT). At the same time, these institutions encourage the use of stored information about patients and molecular processes to reduce the number of RCTs whenever possible. The combination of molecular data linked to clinical phenotypes can render meaningful patterns. When molecular data are available, it is crucial to as certain the relationship between the known clinical phenotypes and the molecular profile of patients. Extracting meaningful clinical patterns from molecular information

can be beneficial. However, obtaining molecular data in RCTs may present practical challenges. These challenges may arise due to difficulties in recruiting patients, particularly for rare diseases. Additionally, RCTs may not always reflect observable results out of controlled settings, necessitating further lengthy observational studies (Andre et al., 2020). 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) (Food and Administration, 2018) and the European Medicines Agency (EMA) (Cave et al., 2019) 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 (Food and Administration)), as well as consistency in data handling and analysis methods (Thorlund et al., 2020), were clear requirements for its acceptance for regulatory purposes (Cave et al., 2019; Zura et al., 2021).

In 1970, Francis Crick presented the Central Dogma of Biology (Crick, 1970).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 (Shapiro, 2009). Therefore, the analysis of an organism's genome and transcriptome sheds light on molecular events underlying human biology and disease (Melé et al., 2015), 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 Microarrays (MA) (1997) (Velculescu et al., 1997) and RNA Sequencing (RNA-Seq) (2006) (Bainbridge et al., 2006) allow the characterisation 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 (Fu et al., 2009) and dynamic range[1] (Zhao et al., 2014) and is particularly suited for *de novo* sequencing since it does not require reference transcripts for probe design (Wang et al., 2009). On the other hand, MA shows higher throughput, is less labour intensive (Minnier et al., 2018) and, due to being an older technology, presents more currently available data (Thompson et al., 2016) which is a key aspect when dealing with big data studies (Belle et al., 2015; Roski et al., 2014). Gene expression values seem to correlate in both techniques (Zhao et al., 2014; Chen et al., 2017; Nookaew et al., 2012), 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 (Park et al., 2021). 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 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 (Stephens et al., 2015). Several scientific works were published exploring the usage of transcriptomics datasets as RWD sources (Arisi et al., 2019; Kato et al., 2020; Fernandes et al., 2021). Table 1 lists four databases containing transcriptomics experimental data. The Gene Expression Omnibus (GEO) (Barrett et al., 2013) 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

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

correct, normalise and summarise transcriptomics data (Chen and Zhou, 2016) are being used, with no consensus. To avoid this issue, most current protocols for data integration involve the usage of raw data (Carey et al., 2018; Junet et al., 2021) and samples from the same experiment (Koeppen et al., 2017). As last resort, the usage of samples involving the same techniques, machines and processing methods is also commonly used (e.g., GEO datasets (Barrett et al., 2013)). 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 (Chen et al., 2019; Khomtchouk et al., 2018) grammar and sentences, hindering its usage in experiments relying on BigData given their lack of machine interpretability (Hadley et al., 2017). Nonetheless, if properly processed and transformed, biological datasets mined from GEO can be used in its fullest and 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	(Barrett et al., 2013)
ArrayExpress	1,309,771	27,068	(Brazma et al., 2003)
Japanese Genotype-phenotype Archive (JGA)	278,601	NA	(Fukuda et al., 2021)
The European Genome-phenome Archive (EGA)	NA	4,432	(Lappalainen et al., 2015)

This paper presents a novel method for analysing GEO MA studies to obtain clinical data. Given the variability in HT data, we developed a technique called MicroArray Cross-plAtfoRm pOst-prOcessiNg (MACAROON) to correct and standardise MA data from different experiments and processing methods. First, we used text mining to standardise clinical labels in the HT repository and identify only patient data, excluding non-human and immortalised cell line data. This information is later analysed alongside language processing techniques to build ML models that clarify original data transformations and address poor metadata issues. Next, we propose a protocol that uses HK genes' expression values to adjust for different normalisation techniques applied to MA datasets before they were uploaded to GEO. This approach allows researchers to combine MA profiles from multiple sources, regardless of the samples' pre-processing steps. Additionally, text mining standardises metadata by classifying samples based on health status and demographics, creating a source of RWD. This method allows to link this RWD with molecular patterns by exploring the transcriptomic data from HT experiments. With this effort, we aim at easing access to RWD, which is currently a main limitation of pharmaceutical companies in this kind of studies (Wallach et al., 2021).

2 Results and Discussion

The foremost objective of this work is to exploit the GEO database (Barrett et al., 2013) as a source of molecular and phenotypical data (RWD). 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 presented herein, 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).

Fig. 1 The overall pipeline of the methods applied in this work. (A) Collection and processing of MA data. (B) Collection and processing of phenotypical data 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 uniformise the orders of magnitude in the samples' data. Then, we apply a GEO-wide normalisation 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 to label samples for the presence of terms related to data processing, like the application of different logarithms, common MA normalisation 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 (combined into ensemble models) – that we will call "$ML_{processing}$" moving forward. These terms were picked to maximise the distinguishability of "applicable" from "non-applicable" classes (when a processing method was applied or when that specific method was not applied, respectively) – 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) normalisation. 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. 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 datasets mentioned above were used to train $ML_{processing}$ to infer if each determined processing technique was applied to the sample or not. 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. 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 1). Yet, in Normalisation, 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. 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 25-k folds CV performance metrics of each model included in the three ensembles. The $ML_{processing}$ were, then, trained sequentially with k-1 parts and tested using the remaining set (Refaeilzadeh et al., 2009). 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 complete dataset. 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 (Koeppen et al., 2017; Chen et al., 2019; Khomtchouk et al., 2018; Hadley et al., 2017). 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.

$$\text{AAC (\%)} = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{1}$$

$$\text{PRE (\%)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \tag{2}$$

$$\text{SNS (\%)} = \frac{\text{TN}}{\text{FP} + \text{FN}} \times 100 \tag{3}$$

$$\text{SPC (\%)} = \frac{\text{TN}}{\text{FP} + \text{TN}} \times 100 \tag{4}$$

*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

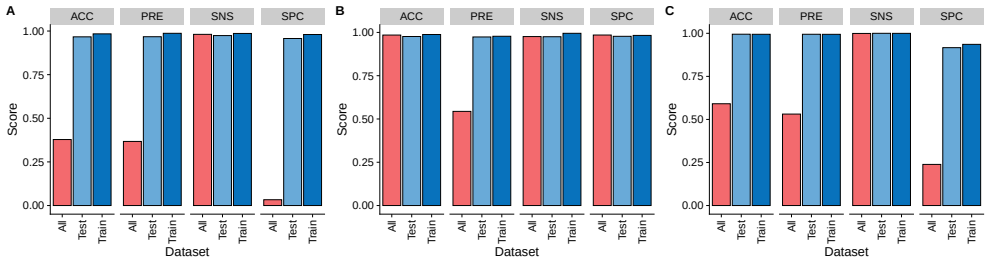

Fig. 2 Global test metrics of the various ensemble models in $\mathbf{ML_{processing}}$. (A) Log2, (B) Log10, and (C) Ratio.

Samples where pre-processing metadata was unclear were classified using the $ML_{processing}$ models. Classification results were used to revert all samples to the original state by reverting the specific Log operation (see Methods section for detail). Given that $Log(x)$ does not exist if x is negative, and to avoid approximation of imaginary numbers, values were restored to their original form. In other words, we have reverted the *Log* transformation in all samples (see 16) instead of uniformising the data by applying *Log*.

Algorithm 1 Composition of ensemble models for the classification of previous processing.

for MA_sample in GEO **do**
  if $log_processing$ is known **then**
    $log_processing \leftarrow \log_processing(X)$
  else
    $log_processing \leftarrow ML_{processing}(MA_sample_{raw})$
  end if
  if is_raw is True **then**
    $MA_sample_{uniform} \leftarrow MA_sample_{raw}$ ▷ Do nothing
  else if $log_processing$ is log2 and not log10 **then**
    $MA_sample_{uniform} \leftarrow 2^{MA_sample_{raw}}$
  else if $log_processing$ is log10 and not log2 **then**
    $MA_sample_{uniform} \leftarrow 10^{MA_sample_{raw}}$
  else
    $MA_sample_{uniform} \leftarrow MA_sample_{raw}$ ▷ Do nothing
  end if
end for

House Keeping-based regression for normalisation

After the first stage of this method – aimed at cleaning and uniformising original processing techniques, we designed a normalisation method based on HK gene expression. The median expression of a set of 40 HK genes selected for their least cross-tissue variability as found in the literature (Eisenberg and Levanon, 2013) was used as reference. For 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. MA normalisation protocols reliant on regression techniques usually use the LOWESS algorithm (Zahurak et al., 2007). However, this solution could not be used here case as it does not allow for extrapolation. Given that HK genes are rarely under/overexpressed, extrapolation was a requirement.

For benchmarking, a test dataset containing MA data from prostate cancer samples and healthy controls was extracted. This dataset included 394 samples from 63 GEO experiments (as available in Supplementary Table 3). 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 (Vafaee et al., 2019), the mean intensity should show a tendency towards a constant value. This tendency is visible from the raw datasets (figure 3A) to the data after previous pre-processing correction (3B), and then to the data subjected to HK regression (3C, and 3D). The effect is also clear on the MA-plots (bottom of figure 3c), where M is defined by equation 5 and A is defined by equation **??**. With the same premise discussed above, the expression of most genes stays generally constant across samples. Consequently, the M value corresponding to the great majority of genes should stay constant along with A if the data is well normalised. 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 localised 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 (Fu et al., 2009).

$$M = \log_2 \left(\frac{Test_i}{Control_i} \right) \tag{5}$$

$$A = \frac{1}{2} \log_2 (Test_i \times Control_i) \tag{6}$$

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

MACAROON compared against YuGene

As further validation of MACAROON, we compared its output against YuGene (Cao et al., 2014): to our knowledge, the only available method that, like the method presented in this paper, is able to make MA normalisation at the sample level. Figures 4A, B and C present a sequence of bar plots describing the intensity values of the GEO series GSE102124 (Sowalsky et al., 2018) (one of the samples available at the PCa dataset used above in this work) – corresponding to 22 MA samples. The subplots 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 log_2

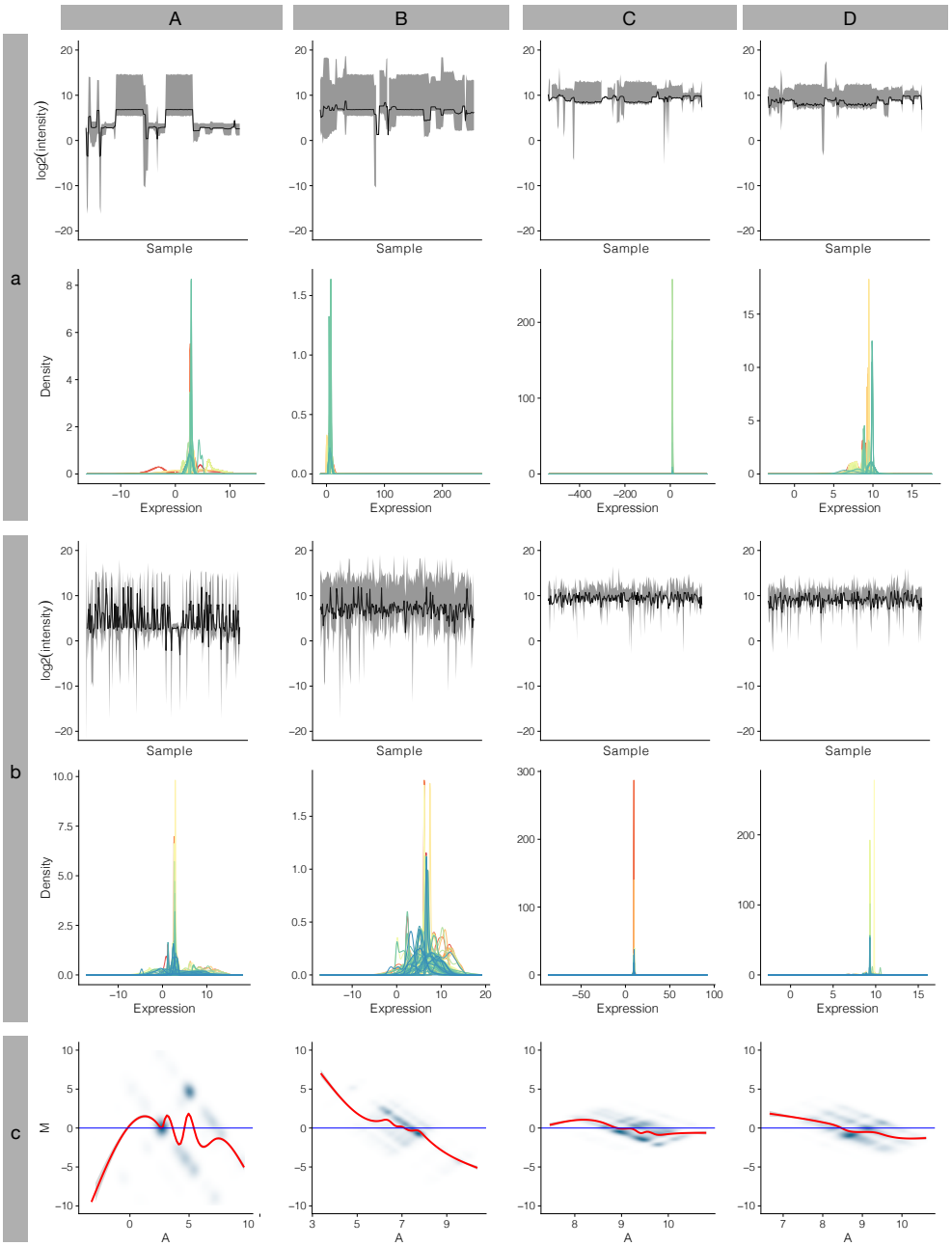

Fig. 3 Log-expression data distribution across the various stages of MACAROON. A, B, C and D refer to the datasets produced along the process: A. Raw unprocessed data; B. After reverting preprocessing; C. After spline regression; and D. After polynomial regression. (a), (b), and (c) concern the samples 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.

scale, the application of MACAROON provides better equivalence to 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' medians, which also impacts 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 lead to datasets with many missing values. Our normalisation 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.

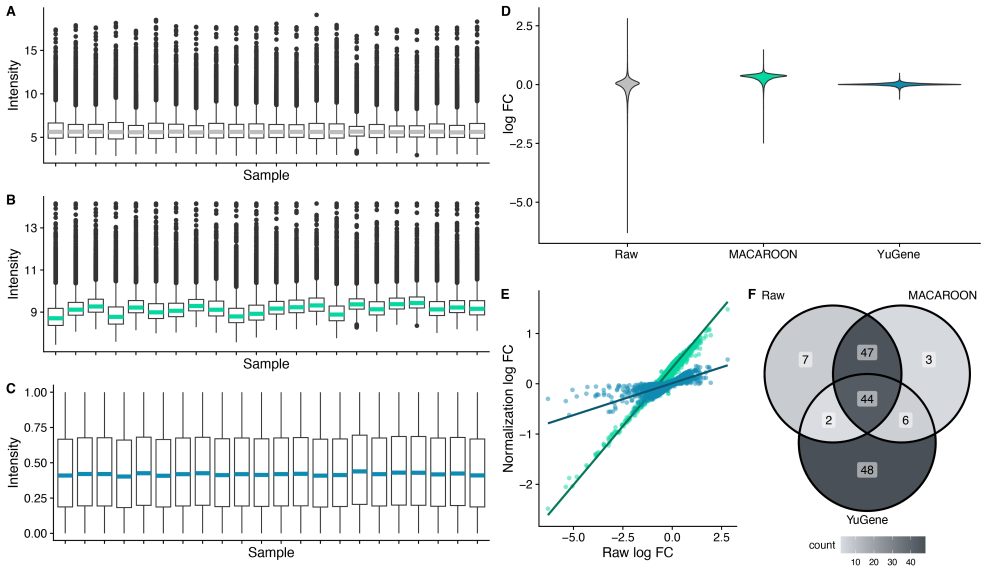

Fig. 4 Direct comparison between methods. In the left column, we show the intensity per sample. A corresponds to the log_2 raw data; B, using MACAROON; and C, using YuGene (Cao et al., 2014). In the right column, we display a comparison of the results of a Differential Expression analysis. In D, we show a dispersion of the log Fold Change values; in E, a scatter plot correlating the fold change values per gene using MACAROON and YuGene vs. the values calculated using the log_2 Raw dataset; in F, a Venn diagram portraying common (ttop-50 and bottom-50 DE genes in log FC) differentially expressed genes depending on the method used. When applicable, the elements in grey correspond to the log_2 Raw dataset, in green to MACAROON, and in blue to YuGene.

As means of further comparing the three methods, we ran a DE Analysis with data processed using the different methods (log_2 Raw, MACAROON, YuGene). In figure 4D, we represent the results' space (density of log Fold-change (FC)) of the DE according to the dataset. 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 the correlation between log Fold-change (FC) results (figure 4E), as well as overlapping differentially expressed genes (figure 4F), MACAROON shows greater similarity to the original results (91% matching) in comparison to YuGene (46%), suggesting better biological interpretability.

Case Study: Differential Analysis Experiment

The integrated PCa dataset presented earlier (available in Supplementary Table 3) was used again to present the method's usability in the real world, and test it by comparing possible biological conclusions, through a 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 figure 5A and B via volcano plots, the full list of results is available in the Supplementary Material 4). The analysis of these materials elucidates that both regression methods are equivalent which suggests that the spline regression detected a linear pattern in these MA data. This conclusion goes in line with descriptions in the literature. According to the characterisation provided by David, E. (Edwards, 2003), 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); these differences are significant in polynomial regression, but not in the case of spline.

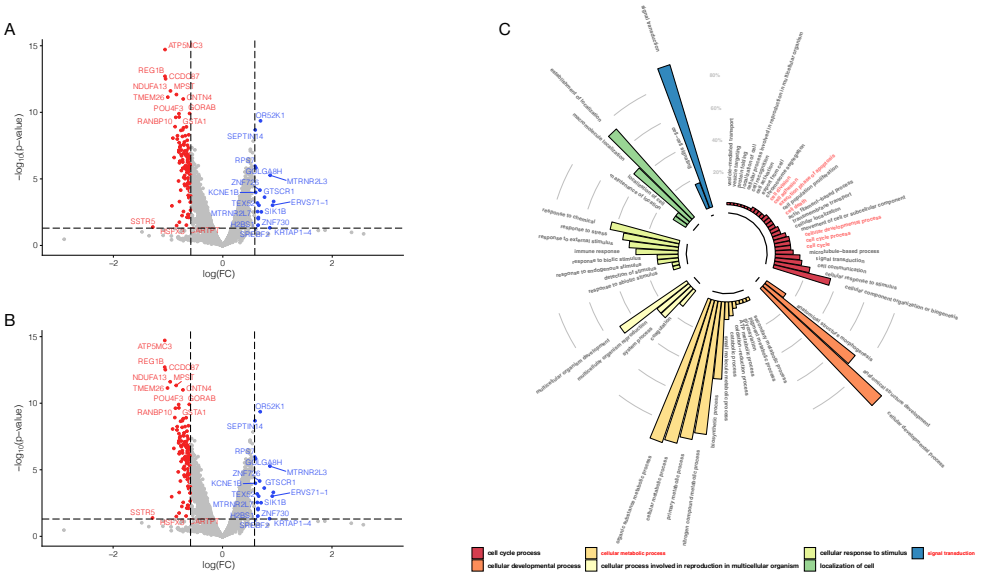

Fig. 5 DE analysis results using spline and polynomial regressions. A, B. Volcano plots of the DE using both regression techniques. C. Enrichment analysis results using the polynomial dataset (where biological processes found in other works in literature are highlighted in red). In A and B, the dots 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 biological processes, while each column refers to the second rank; the axis reflects the percentage of DE genes in the list associated with a particular annotation term.

$$i_{ps} \sim k_s \cdot \alpha_p \cdot \theta_p \tag{7}$$

*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 (GO) (Carbon et al., 2021) biological processes using PANTHER (Mi et al., 2021). 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, biological processes found in other publications on Prostate Cancer are highlighted in red. (Yang et al., 2014) reported differences in biological adhesion, while (Fan et al., 2018) described changes in cell cycle control, cell death regulation, cell development, and focal adhesion. (Sun et al., 2019) and (Verma et al., 2019) reported alterations in the hippo signalling pathway, negative regulation of growth, and multiple metabolic progression.

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 (Montes et al., 2018; Przemska-Kosicka et al., 2018), biological sex (Basoglu and Tasbakan, 2018; Gabler et al., 2012) or ethnical background (Morris et al., 2019; Kopel et al., 2020). Additionally, in the case of drug studies, information about body weight is also crucial for the analysis/simulation of pharmacokinetics (Hebbes and Thompson, 2018; Morrish et al., 2011; Smit et al., 2018) and to better understand other unexpected effects (e.g. "BMI paradox" (Tittl et al., 2018; Donini et al., 2020) – in which, patients with extreme weight seem to exhibit 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 organise them. Figure 6 presents a graphical description of the obtained GEO patient populations. 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, B, and C, 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.

Mining demographical metadata from GEO samples is not a new subject. Other works have accomplished this task either by manual curation (STARGEO (Hadley et al., 2017), and OMiCC (Shah et al., 2016)) or by applying AI and NLP techniques (MetaSRA (Bernstein et al., 2017) and ALE (Giles et al., 2017)). 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., 2019) in their review of the subject. 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

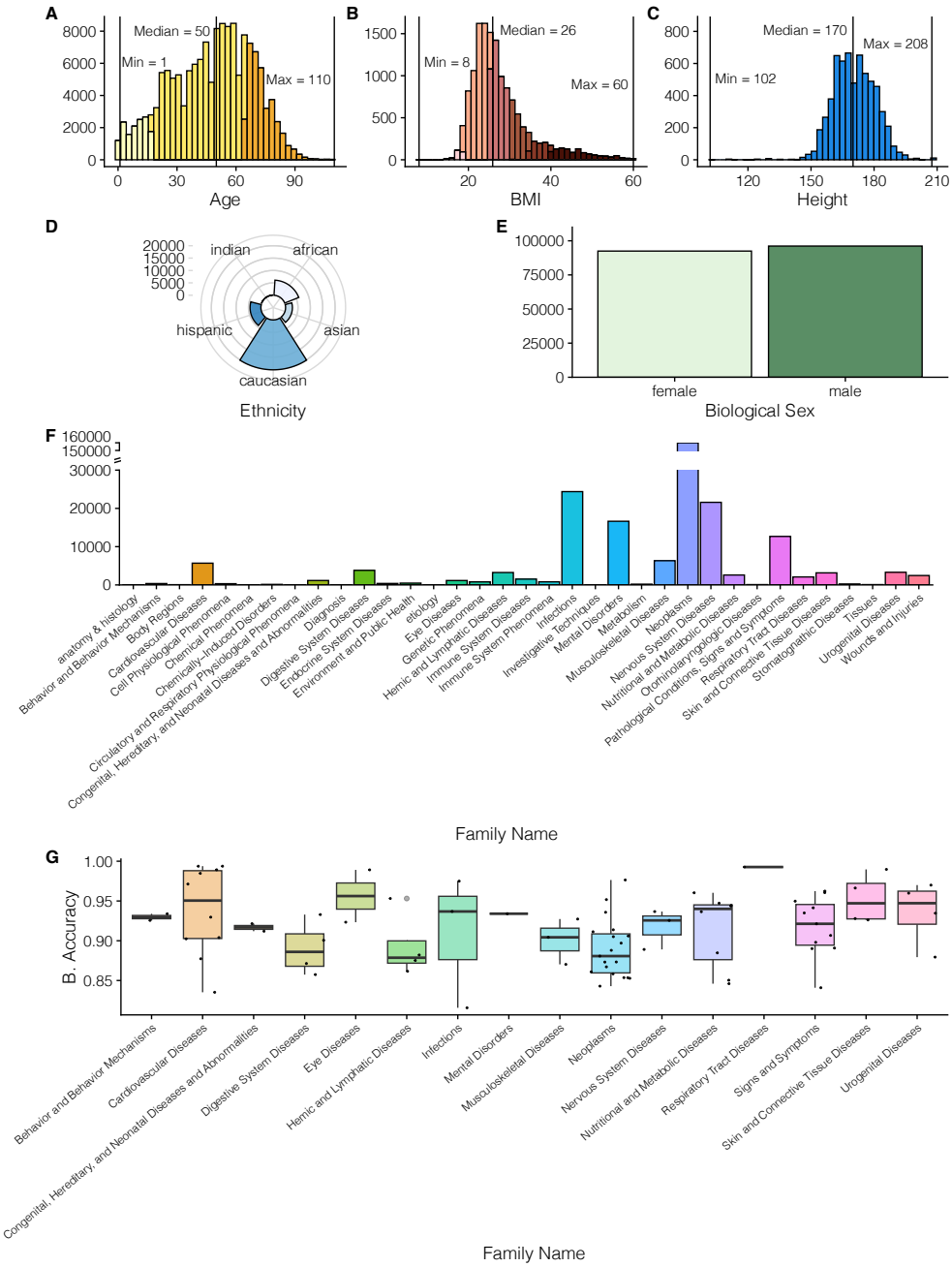

Fig. 6 Demographic description of MA samples in GEO. A, Age distribution of samples (n=152,237). Colours represent various age groupings: light yellow – childhood; yellow – adulthood; orange – elderhood. B, BMI distribution of samples (n=14,243). Different shades of red represent different gradings of GEO: 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.

the accompanying text (present in GEO samples' metadata) was used to assign mentioned medical conditions and their status as healthy, control, or intervention case. To use this RWD repository in a TPMS framework that relies on the Biological Effectors Database (BED) (Jorba et al., 2020), which maps medical conditions to their protein effectors, we mined metadata for the medical terms in BED (300 conditions at the time of analysis). BED links biological processes (adverse drug events, drug indications, medical conditions) with their associated proteins, based on literature reviews that list genes and proteins related to specific conditions and their roles in disease pathways. Extracted pathological attributes in GEO samples were mapped to the BED condition terms. 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). However, from the 300 total terms only a portion had corresponding samples in GEO, leading to the final total 162 tagged medical conditions. Regardless, linking these terms to BED permitted associating each medical condition and consequently each sample in GEO to literature-derived molecular descriptions of disease. Figure 6F describes the overall distribution of GEO patients across Medical Subject Headings (MeSH) (for Biotechnology Information and of Medicine, 2021) 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 (e.g., cancer, neurological and psychological diseases), and not to those more common in societies.

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 to be predicted. Features were defined by BED information on disease-related proteins/-genes. 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 corresponding to the same BED medical condition to be able to counteract the large number of missing probes. A global depiction of the CV Ballanced Accuracy (BACC) (Equation 8) metrics is displayed in figure 6G. Here, given the large number of BED conditions, different pathologies are grouped into MeSH condition groups – as in Figure 6F; for further detail, extensive results are available in Supplementary Table 7. 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. Across all available BED diseases, $ML_{clinical}$ models reached at least 81.57% balanced accuracy (average: 91.47%; standard deviation: 0.044). 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 (in addition to the DE analysis presented and discussed in previous sections). Furthermore, these models will also be used for filling the gaps in missing metadata, as well as studying possible comorbidities.

$$\text{BACC} = \frac{SNS + SPC}{2} \tag{8}$$

3 Conclusion

In this work, we presented 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). Our approach begins with an acknowledgment of the predominance of processed data within GEO. Our methodology revolves around the integration of MA data from diverse sources, encompassing various MA platforms, into a unified dataset. Our emphasis lies in the ability to incorporate samples from past experiments, particularly in scenarios where raw intensity values are absent, and investigators only supply processed and normalised files. Through this concerted effort, we aim to streamline data management and improve the comprehensiveness of MA datasets within the GEO repository. The different processing techniques for the normalisation of MA data, and the fact that often supplementary metadata lacks precise information, negatively influence the standardisation 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 normalisation 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 biological mechanisms 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 guidance of typical information reported in 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 *in silico* experiments and explore an even more diverse set of patients.

4 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 (Barrett et al., 2013). 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 steps. This procedure was performed with Python 3.9.0 using the package Beautiful Soup 4 (Richardson, 2007). 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 mapped to their corresponding gene id. 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 normalisation)

Composition of datasets

A series of expression-related features were extracted from the MA datasets mentioned in the previous steps. Statistical descriptions such as: 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 (Butte et al., 2001) and are universally expressed in all tissues and cell types (Zhu et al., 2008). Given their known variation in the expression levels in some or conditions (Greer et al., 2010), we ranked known HK genes according to the standard deviation of their expression across multiple tissues,

as published by (Eisenberg and Levanon, 2013). Additionally, as means of maximizing consensus, 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) (Zhang et al., 2015). 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 pre-processing classification.

Uniprot	**Gene Name**	**Protein Function**	$\sigma(\log_2((RPKM))$
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

Table 3: **List of HK genes included on the datasets used in this work organised by standard deviation of expression across multiple tissues.** *(continued)*

Uniprot	Gene Name	Protein Function	$\sigma(\log_2((RPKM))$
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

Table 3: **List of HK genes included on the datasets used in this work organised by standard deviation of expression across multiple tissues.** *(continued)*

Uniprot	Gene Name	Protein Function	$\sigma(\log_2((RPKM))$
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

Table 3: **List of HK genes included on the datasets used in this work organised by standard deviation of expression across multiple tissues.** *(continued)*

Uniprot	Gene Name	Protein Function	$\sigma(\log_2((RPKM))$
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

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

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 normalisation, "quantile", "ratio", "raw" ; and "rma". The resulting dataset included samples with one or more "hot words", as portrayed in table 4. A random selection of 5000 samples, for each of the 6 groups defined 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.

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: Decision tree (Breiman et al., 2017), Elastic net (Zou and Hastie, 2005), Elastic net with threshold optimization (Boughorbel et al., 2017), Generalized linear model (Madsen and Thyregod, 2010), Linear classifier (Fukunaga, 1990), Multi-layer perceptron (Haykin, 1998), Naive Bayes Classifier (Russell and Norvig, 2010), Quadratic classifier (Haykin, 1998), and Support Vector Machines (Ben-Hur et al., 2008). 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) (Inc., 2017) and its "Statistics and Machine Learning Toolbox" (Henson and Cetto, 2005). 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 (supplementary figure 1).

Uniformization of processing techniques

Samples were classified by $ML_{processing}$ models. In cases where the original pre-processing annotation was unavailable or unclear, expression values were transformed according to classification results. The transformations applied are described in equations 9 and 10.

Where log2 is detected,

$$x^{'} = 2^{x} \tag{9}$$

And, where log10 is detected,

$$x^{'} = 10^{x} \tag{10}$$

Cross-sample normalisation method

After undoing logarithm transformations, we applied a normalisation technique to uniformize 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 applied a normalisation 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 normalisation 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 shows 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 (Inc., 2017) and (2) polynomial regressors (Inc., 2017) – 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 (Cao et al., 2014) as means of evaluating its performance in comparison to an original MA dataset (GEO series GSE102124 (Sowalsky et al., 2018) – 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. The analysis was conducted in R using the package *limma* v. 3.19 (Ritchie et al., 2015). Overlap analysis was performed with the top-50 and bottom-50 genes from each dataset in respect to log FC.

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 (Dekking et al., 2005)). 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 with Bonferroni correction. DEGs were analysed for Gene Ontology BP (Carbon et al., 2021) enrichment using PANTHER (Mi et al., 2021).

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 (for Biotechnology Information and of Medicine, 2021) using MetaMap (Bodenreider, 2004) – a tool to map biomedical terms referred to in raw text to controlled terms in the Unified Medical Language System (UMLS) and corresponding Condition ID (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 (Gremse et al., 2011).

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).

Supplementary information. This article has accompanying supplementary files.

Declarations

- 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 Sklodowska-Curie grant agreement No 860303 (proEVLifeCycle) and 859962 (SECRET), respectively.
- Conflict of interest/Competing interests: None declared.

- 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.
- Data availability: 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.
- Materials availability: Not applicable.
- Code availability: Not applicable.
- Author contribution: PMF, methodology, software, validation, formal analysis, investigation, resources, Writing – original draft, Writing – review & editing, visualisation. 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.

References

Jayatunga MKP, Xie W, Ruder L, Schulze U, Meier C. AI in small-molecule drug discovery: a coming wave? Nature Reviews Drug Discovery. 2022 3;21:175–176. https://doi.org/10.1038/D41573-022-00025-1.

Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discovery Today. 2021 1;26:80–93. https://doi.org/10.1016/J.DRUDIS.2020.10.010.

Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Research. 2017 1;45:D985–D994. https://doi.org/10.1093/NAR/GKW1055.

Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nature Machine Intelligence 2020 2:10. 2020 10;2:573–584. https://doi.org/10.1038/s42256-020-00236-4.

Rodriguez S, Hug C, Todorov P, Moret N, Boswell SA, Evans K, et al. Machine learning identifies candidates for drug repurposing in Alzheimer's disease. Nature Communications 2021 12:1. 2021 2;12:1–13. https://doi.org/10.1038/s41467-021-21330-0.

Rashid MBMA, Toh TB, Hooi L, Silva A, Zhang Y, Tan PF, et al. Optimizing drug combinations against multiple myeloma using a quadratic phenotypic optimization

platform (QPOP). Science Translational Medicine. 2018 8;10. https://doi.org/10.1126/SCITRANSLMED.AAN0941/SUPPL_FILE/AAN0941_TABLE_S19.ZIP.

Kolla L, Gruber FK, Khalid O, Hill C, Parikh RB. The case for AI-driven cancer clinical trials – The efficacy arm in silico. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2021 8;1876:188572. https://doi.org/10.1016/J.BBCAN.2021.188572.

Collins GS, Dhiman P, Navarro CLA, Ma J, Hooft L, Reitsma JB, et al. 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. 2021 7;11:e048008. https://doi.org/10.1136/BMJOPEN-2020-048008.

Andre EB, Reynolds R, Caubel P, Azoulay L, Dreyer NA. Trial designs using real-world data: The changing landscape of the regulatory approval process. Pharmacoepidemiology and Drug Safety. 2020 10;29:1201–1212. https://doi.org/10.1002/PDS.4932.

Food, Administration D. Framework for FDA's Real-World Evidence Program. 2018;.

Cave A, Kurz X, Arlett P. Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe. Clinical Pharmacology Therapeutics. 2019 7;106:36–39. https://doi.org/10.1002/CPT.1426.

Food, Administration D.: Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices — FDA. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices.

Thorlund K, Dron L, Park JJH, Mills EJ. Synthetic and external controls in clinical trials – A primer for researchers. Clinical Epidemiology. 2020;12:457–467. https://doi.org/10.2147/CLEP.S242097.

Zura R, Irwin DE, Mack CD, Aldridge ML, Mackowiak JI. Real-World Evidence: A Primer. Journal of orthopaedic trauma. 2021 3;35. https://doi.org/10.1097/BOT.0000000000002037.

Crick F. Central Dogma of Molecular Biology. Nature 1970 227:5258. 1970;227:561–563. https://doi.org/10.1038/227561a0.

Shapiro JA. Revisiting the central dogma in the 21st century. Annals of the New York Academy of Sciences. 2009 10;1178:6–28. https://doi.org/10.1111/J.1749-6632.2009.04990.X.

Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. Science. 2015 5;348:660–665. https://doi.org/10.1126/SCIENCE.AAA0355/SUPPL_FILE/MELE.SM.PDF.

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, et al. Characterization of the Yeast Transcriptome. Cell. 1997 1;88:243–251. https://doi.org/10.1016/S0092-8674(00)81845-0.

Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, et al. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics. 2006 9;7:1–11. https://doi.org/10.1186/1471-2164-7-246/TABLES/5.

Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, et al. Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics. 2009 4;10:1–9. https://doi.org/10.1186/1471-2164-10-161/FIGURES/3.

Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLOS ONE. 2014 1;9:e78644. https://doi.org/10.1371/JOURNAL.PONE.0078644.

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics. 2009 1;10:57. https://doi.org/10.1038/NRG2484.

Minnier J, Pennock ND, Guo Q, Schedin P, Harrington CA. RNA-seq and expression arrays: Selection guidelines for genome-wide expression profiling. Methods in Molecular Biology. 2018;1783:7–33. https://doi.org/10.1007/978-1-4939-7834-2_2.

Thompson JA, Tan J, Greene CS. Cross-platform normalization of microarray and RNA-seq data for machine learning applications. PeerJ. 2016 1;2016:e1621. https://doi.org/10.7717/PEERJ.1621/SUPP-5.

Belle A, Thiagarajan R, Soroushmehr SMR, Navidi F, Beard DA, Najarian K. Big data analytics in healthcare. BioMed Research International. 2015;2015. https://doi.org/10.1155/2015/370194.

Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: Opportunities and policy implications. Health Affairs. 2014 8;33:1115–1122. https://doi.org/10.1377/HLTHAFF.2014.0147/ASSET/IMAGES/LARGE/2014.0147FIGEX1.JPEG.

Chen L, Sun F, Yang X, Jin Y, Shi M, Wang L, et al. Correlation between RNA-Seq and microarrays results using TCGA data. Gene. 2017 9;628:200–204. https://doi.org/10.1016/J.GENE.2017.07.056.

Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, et al. 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 Research. 2012 11;40:10084–10097. https://doi.org/10.1093/NAR/GKS804.

Park SH, Choi J, Byeon JS. Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence. Korean Journal of Radiology. 2021 3;22:442–453. https://doi.org/10.3348/KJR.2021.0048.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLOS Biology. 2015 7;13:e1002195. https://doi.org/10.1371/journal.pbio.1002195.

Arisi I, Dronofrio M, Brandi R, Sonnessa M, Campanelli A, Florio R, et al. Mining clinical and laboratory data of neurodegenerative diseases by Machine Learning: Transcriptomic biomarkers. Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018. 2019 1;p. 2735–2737. https://doi.org/10.1109/BIBM.2018.8621072.

Kato S, Kim KH, Lim HJ, Boichard A, Nikanjam M, Weihe E, et al. Real-world data from a molecular tumor board demonstrates improved outcomes with a precision N-of-One strategy. Nature Communications 2020 11:1. 2020 10;11:1–9. https://doi.org/10.1038/s41467-020-18613-3.

Fernandes LE, Epstein CG, Bobe AM, Bell JSK, Stumpe MC, Salazar ME, et al. Real-world Evidence of Diagnostic Testing and Treatment Patterns in US Patients With Breast Cancer With Implications for Treatment Biomarkers From RNA Sequencing Data. Clinical Breast Cancer. 2021 8;21:e340–e361. https://doi.org/10.1016/J.CLBC.2020.11.012.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: Archive for functional genomics data sets - Update. Nucleic Acids Research.

2013;https://doi.org/10.1093/nar/gks1193.

Chen N, Zhou Q. The evolving gleason grading system. Chinese Journal of Cancer Research. 2016 2;28:58–64. https://doi.org/10.3978/J.ISSN.1000-9604.2016.02.04.

Carey M, Ramírez JC, Wu S, Wu H. A big data pipeline: Identifying dynamic gene regulatory networks from time-course Gene Expression Omnibus data with applications to influenza infection. Statistical Methods in Medical Research. 2018 7;27:1930–1955. https://doi.org/10.1177/0962280217746719.

Junet V, Farrés J, Mas JM, Daura X. CuBlock: a cross-platform normalization method for gene-expression microarrays. Bioinformatics. 2021 8;37:2365–2373. https://doi.org/10.1093/BIOINFORMATICS/BTAB105.

Koeppen K, Stanton BA, Hampton TH. ScanGEO: parallel mining of high-throughput gene expression data. Bioinformatics. 2017 11;33:3500–3501. https://doi.org/10.1093/BIOINFORMATICS/BTX452.

Chen BY, Sung CWH, Chen C, Cheng CM, Lin DPC, Huang CT, et al. Advances in exosomes technology. Clinica Chimica Acta. 2019 6;493:14–19. https://doi.org/10.1016/J.CCA.2019.02.021.

Khomtchouk BB, Dyomkin V, Vand KA, Assimes T, Gozani O. Biochat: a database for natural language processing of Gene Expression Omnibus data. bioRxiv. 2018 11;p. 480020. https://doi.org/10.1101/480020.

Hadley D, Pan J, El-Sayed O, Aljabban J, Aljabban I, Azad TD, et al. Precision annotation of digital samples in NCBI's gene expression omnibus. Scientific Data 2017 4:1. 2017 9;4:1–11. https://doi.org/10.1038/sdata.2017.125.

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Research. 2003 1;31:68–71. https://doi.org/10.1093/NAR/GKG091.

Fukuda A, Kodama Y, Mashima J, Fujisawa T, Ogasawara O. DDBJ update: streamlining submission and access of human data. Nucleic Acids Research. 2021 1;49:D71–D75. https://doi.org/10.1093/NAR/GKAA982.

Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, ur Rehman S, et al. The European Genome-phenome Archive of human data consented for biomedical research. Nature Genetics 2015 47:7. 2015 6;47:692–695. https://doi.org/10.1038/ng.3312.

Wallach JD, Zhang AD, Skydel JJ, Bartlett VL, Dhruva SS, Shah ND, et al. Feasibility of Using Real-world Data to Emulate Postapproval Confirmatory Clinical Trials of Therapeutic Agents Granted US Food and Drug Administration Accelerated Approval. JAMA Network Open. 2021 11;4:e2133667–e2133667. https://doi.org/10.1001/JAMANETWORKOPEN.2021.33667.

Refaeilzadeh P, Tang L, Liu H. Cross-Validation. Encyclopedia of Database Systems. 2009;p. 532–538. https://doi.org/10.1007/978-0-387-39940-9_565.

Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends in Genetics. 2013 10;29:569–574. https://doi.org/10.1016/J.TIG.2013.05.010.

Zahurak M, Parmigiani G, Yu W, Scharpf RB, Berman D, Schaeffer E, et al. Preprocessing Agilent microarray data. BMC Bioinformatics. 2007 5;8:1–13. https://doi.org/10.1186/1471-2105-8-142/FIGURES/10.

Vafaee F, Dashti H, Alinejad-Rokny H. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Transcriptomic Data Normalization. vol. 2. Academic Press;

2019. p. 364–371.

Cao KAL, Rohart F, McHugh L, Korn O, Wells CA. YuGene: A simple approach to scale gene expression data derived from different platforms for integrated analyses. Genomics. 2014 4;103:239–251. https://doi.org/10.1016/J.YGENO.2014.03.001.

Sowalsky AG, Ye H, Bhasin M, Allen EMV, Loda M, Lis RT, et al. Neoadjuvant-Intensive Androgen Deprivation Therapy Selects for Prostate Tumor Foci with Diverse Subclonal Oncogenic Alterations. Cancer research. 2018 8;78:4716–4730. https://doi.org/10.1158/0008-5472.CAN-18-0610.

Edwards D. Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics. 2003 5;19:825–833. https://doi.org/10.1093/BIOINFORMATICS/BTG083.

Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, et al. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research. 2021 1;49:D325–D334. https://doi.org/10.1093/NAR/GKAA1113.

Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Research. 2021 1;49:D394–D403. https://doi.org/10.1093/NAR/GKAA1106.

Yang J, Chen L, Kong X, Huang T, Cai YD. Analysis of Tumor Suppressor Genes Based on Gene Ontology and the KEGG Pathway. PLOS ONE. 2014 9;9:e107202. https://doi.org/10.1371/JOURNAL.PONE.0107202.

Fan S, Liang Z, Gao Z, Pan Z, Han S, Liu X, et al. Identification of the key genes and pathways in prostate cancer. Oncology Letters. 2018 11;16:6663–6669. https://doi.org/10.3892/OL.2018.9491/HTML.

Sun J, Li S, Wang F, Fan C, Wang J. Identification of key pathways and genes in pten mutation prostate cancer by bioinformatics analysis. BMC Medical Genetics. 2019 12;20:1–9. https://doi.org/10.1186/S12881-019-0923-7/FIGURES/6.

Verma S, Shukla S, Pandey M, MacLennan GT, Gupta S. Differentially expressed genes and molecular pathways in an autochthonous mouse prostate cancer model. Frontiers in Genetics. 2019;10:235. https://doi.org/10.3389/FGENE.2019.00235/BIBTEX.

Montes J, McDermott MP, Mirek E, Mazzone ES, Main M, Glanzman AM, et al. Ambulatory function in spinal muscular atrophy: Age-related patterns of progression. PLOS ONE. 2018 6;13:e0199657. https://doi.org/10.1371/JOURNAL.PONE.0199657.

Przemska-Kosicka A, Childs CE, Maidens C, Dong H, Todd S, Gosney MA, et al. Age-related changes in the natural killer cell response to seasonal influenza vaccination are not influenced by a synbiotic: A randomised controlled trial. Frontiers in Immunology. 2018 3;8:591. https://doi.org/10.3389/FIMMU.2018.00591/BIBTEX.

Basoglu OK, Tasbakan MS. Gender differences in clinical and polysomnographic features of obstructive sleep apnea: a clinical study of 2827 patients. Sleep and Breathing. 2018 3;22:241–249. https://doi.org/10.1007/S11325-017-1482-9/TABLES/4.

Gabler NB, French B, Strom BL, Liu Z, Palevsky HI, Taichman DB, et al. Race and Sex Differences in Response to Endothelin Receptor Antagonists for Pulmonary Arterial Hypertension. Chest. 2012 1;141:20–26. https://doi.org/10.1378/CHEST.11-0404.

Morris JC, Schindler SE, McCue LM, Moulder KL, Benzinger TLS, Cruchaga C, et al. Assessment of Racial Disparities in Biomarkers for Alzheimer Disease. JAMA Neurology. 2019 3;76:264–273. https://doi.org/10.1001/JAMANEUROL.2018.4249.

Kopel J, Perisetti A, Roghani A, Aziz M, Gajendran M, Goyal H. Racial and Gender-Based Differences in COVID-19. Frontiers in Public Health. 2020 7;8:418. https://doi.org/10.3389/FPUBH.2020.00418/BIBTEX.

Hebbes CP, Thompson JP. Pharmacokinetics of anaesthetic drugs at extremes of body weight. BJA Education. 2018 12;18:364–370. https://doi.org/10.1016/J.BJAE.2018.09.001/ATTACHMENT/5E102D20-4DC4-43D0-8CD6-7AA6C475F22F/MMC1.DOCX.

Morrish GA, Pai MP, Green B. The effects of obesity on drug pharmacokinetics in humans. http://dxdoiorg/101517/174252552011570331. 2011 6;7:697–706. https://doi.org/10.1517/17425255.2011.570331.

Smit C, Hoogd SD, Brüggemann RJM, Knibbe CAJ. Obesity and drug pharmacology: a review of the influence of obesity on pharmacokinetic and pharmacodynamic parameters. https://doiorg/101080/1742525520181440287. 2018 3;14:275–285. https://doi.org/10.1080/17425255.2018.1440287.

Tittl L, Endig S, Marten S, Reitter A, Beyer-Westendorf I, Beyer-Westendorf J. Impact of BMI on clinical outcomes of NOAC therapy in daily care - Results of the prospective Dresden NOAC Registry (NCT01588119). International Journal of Cardiology. 2018 7;262:85–91. https://doi.org/10.1016/J.IJCARD.2018.03.060.

Donini LM, Pinto A, Giusti AM, Lenzi A, Poggiogalle E. Obesity or BMI Paradox? Beneath the Tip of the Iceberg. Frontiers in Nutrition. 2020 5;7:53. https://doi.org/10.3389/FNUT.2020.00053/BIBTEX.

Shah N, Guo Y, Wendelsdorf KV, Lu Y, Sparks R, Tsang JS. A crowdsourcing approach for reusing and meta-analyzing gene expression data. Nature Biotechnology 2016 34:8. 2016 6;34:803–806. https://doi.org/10.1038/nbt.3603.

Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. Bioinformatics. 2017 9;33:2914–2923. https://doi.org/10.1093/BIOINFORMATICS/BTX334.

Giles CB, Brown CA, Ripperger M, Dennis Z, Roopnarinesingh X, Porter H, et al. ALE: Automated label extraction from GEO metadata. BMC Bioinformatics. 2017 12;18:7–16. https://doi.org/10.1186/S12859-017-1888-1/FIGURES/10.

Wang Z, Lachmann A, Ma'ayan A. Mining data and metadata from the gene expression omnibus. Biophysical Reviews. 2019 2;11:103–110. https://doi.org/10.1007/S12551-018-0490-8/FIGURES/2.

Jorba G, Aguirre-Plans J, Junet V, Segú-Vergés C, Ruiz JL, Pujol A, et al. In-silico simulated prototype-patients using TPMS technology to study a potential adverse effect of sacubitril and valsartan. PLOS ONE. 2020 2;15:e0228926. https://doi.org/10.1371/JOURNAL.PONE.0228926.

for Biotechnology Information NC, of Medicine USNL.: MeSH (Medical Subject Headings). Available from: https://www.ncbi.nlm.nih.gov/mesh/.

Richardson L.: Beautiful soup documentation.

Butte AJ, Dzau VJ, Glueck SB. Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". https://doiorg/101152/physiolgenomics20017295. 2001;7:95–96. https://doi.org/10.1152/PHYSIOLGENOMICS.2001.7.2.95.

Zhu J, He F, Hu S, Yu J. On the nature of human housekeeping genes. Trends in Genetics. 2008 10;24:481–484. https://doi.org/10.1016/J.TIG.2008.08.004.

Greer S, Honeywell R, Geletu M, Arulanandam R, Raptis L. Housekeeping genes; expression levels may change with density of cultured cells. Journal of Immunological Methods. 2010 4;355:76–79. https://doi.org/10.1016/J.JIM.2010.02.006.

Zhang Y, Li D, Sun B. Do Housekeeping Genes Exist? PLOS ONE. 2015 5;10:e0123691. https://doi.org/10.1371/JOURNAL.PONE.0123691.

Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Classification and Regression Trees. 2017 1;p. 1–358. https://doi.org/10.1201/9781315139470/CLASSIFICATION-REGRESSION-TREES-LEO-BREIMAN.

Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2005;67:301–320.

Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE. 2017 6;12:e0177678. https://doi.org/10.1371/JOURNAL.PONE.0177678.

Madsen H, Thyregod P. Introduction to general and generalized linear models. Chapman Hall/CRC; 2010. Available from: https://www.taylorfrancis.com/books/mono/10.1201/9781439891148/introduction-general-generalized-linear-models-henrik-madsen-poul-thyregod.

Fukunaga K. Introduction to Statistical Pattern Recognition. Academic Press; 1990.

Haykin S. Neural Networks: A Comprehensive Foundation. 2nd ed. Prentice Hall PTR; 1998.

Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. 3rd ed. Pearson Education, Inc.; 2010.

Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support Vector Machines and Kernels for Computational Biology. PLOS Computational Biology. 2008;4:e1000173. https://doi.org/10.1371/JOURNAL.PCBI.1000173.

Inc TM.: MATLAB: r2017b. The Mathworks, Inc.

Henson R, Cetto L. The MATLAB bioinformatics toolbox. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. 2005 1;https://doi.org/10.1002/047001153X.G409308.

Inc TM.: MATLAB Smoothing Splines. The Mathworks, Inc. Available from: https://www.mathworks.com/help/curvefit/smoothing-splines.html.

Inc TM.: MATLAB Polynomial Models. The Mathworks, Inc. Available from: https://www.mathworks.com/help/curvefit/polynomial.html.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015 4;43:e47–e47. https://doi.org/10.1093/NAR/GKV007.

Dekking FM, Kraaikamp C, Lopuhaä HP, Meester LE. Exploratory data analysis: numerical summaries. 2005;p. 231–243. https://doi.org/10.1007/1-84628-168-7_16.

Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004 1;32:D267–D270. https://doi.org/10.1093/NAR/GKH061.

Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme

sources. Nucleic Acids Research. 2011 1;39:D507–D513. https://doi.org/10.1093/NAR/GKQ968.