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

Introduction:

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:

Samples from patient data were integrated into the patient repository by using a new post-processing technique allowing for the combined usage of samples originating from 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, that we've called MACAROON, 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% enhancement (compared to other methods available). RWD mining was done through a manually curated synonym dictionary allowing for the correct assignment (95.33% median accuracy) of medical conditions.

Conclusion:

Our strategy produced a RWD repository, including molecular information and clinical and demographical RWD. The exploration of these data facilitates shedding light on 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 boost the acceleration timeline of production of pharmaceutical products, therefore, aiding in cost reduction[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]. The usage of AI, and specifically ML algorithms, is dependent on high volumes of data[9]. Nonetheless, regulatory agencies still mostly accept data acquired through RCT. Also, the same regulatory agencies encourage the use of stored information about patients and molecular processes to reduce the number of RCTs when it is possible. When molecular data are available, it is required 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[10], which requires the execution of further lengthy observational studies. The usage of RWD has been recently considered to be a viable alternative to complement clinical conclusions. Despite some divergences in minor details (i.e., specific data source), RWD is defined as data relating to patients' health status that didn't originate in traditional clinical trials. Since 2018, both the US Food and Drug Administration (FDA)[11], and the European Medicines Agency (EMA)[12], laid down their regulatory frameworks from which the usage of RWD could be used to obtain valid clinical evidence. In both cases, accuracy, and reliability

(defined according to the scope of the study[13]) of RWD were clear requirements for its acceptance for regulatory purposes[12,14].

In 1970, Francis Crick presented the Central Dogma of Biology[15]. According to his theory, information that codes proteins in living organisms flow 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[16]. Therefore, the analysis of the information on the genome and transcriptome of an organism sheds light on molecular events underlying human biology and disease[17], producing significant insight into its phenotype, and giving key clues on 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)[18] and RNA Sequencing (RNA-Seq) (2006)[19] 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[20], and dynamic range[21] and is particularly suited for *de novo* sequencing, since it does not require reference transcripts for probe design[22]. On the other hand, MA shows higher throughput, is less labour intensive[23], and, due to being an older technology, presents more currently available data[24] which is a key aspect when dealing with big data problems[25,26]. Furthermore, expression values seem to correlate in both techniques[21,27,28], which validates the usage of both technologies if limitations are considered.

The great volumes of data available today led to the adoption of a culture of sharing and reusing HT-generated data within the scientific community, and to the development of biological databases to publicly host the high volume of new data. After 30 years, the volume of data

hosted yearly in these repositories reaches Terabytes, and it's 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, most entries on the platform only include data already pre-processed by the submitting author. Submitting authors apply a large variety of methods to background correct, normalize and summarize transcriptomics data[31,32]. 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 is often inaccurate information on 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, different scientific works were published exploring the usage of transcriptomics datasets as RWD. 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. Patients 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, explored 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, 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). After that, given the different processing techniques applied to the HT data and its heterogeneity, we developed a novel technique to correct processed MA data obtained from multiple experiments that we've called MACAROON – *MicroArray Cross-plAtfoRm pOst-prOcessiNg*. 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 developed a protocol that takes expression values of House Keeping (HK) genes to readjust the different previous normalization techniques – corresponding to the normalization method applied to MA datasets before uploading to GEO. By applying the technique herein proposed, the researcher is able to combine MA profiles from multiple sources or experiments independently of the pre-processing of samples. 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 the exploration of the transcriptomic patterns underlying the provided HT experiments. With this effort, we aim at easing access to RWD, 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've obtained a total of 1,866,292 MA samples (including gene expression and corresponding metadata). Gene expression data was subjected to a new post-processing method (introduced in this work), while metadata was mined for phenotypical labels (demographics and medical conditions). Regardless, 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 and the gene expression data available was used to generate classifiers aimed at predicting clinical phenotypes excluded from metadata. The whole set of samples was run through these ML models, and, for each sample, new predicted clinical labels were kept – further complementing our RWD repository. In conclusion, each sample in the RWD repository contains the post-processed MA data (transcriptomics data molecularly describing the patient) and the corresponding clinical labels (both original and predicted, including cross-validation accuracy for quality assessment in the latter). 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**.

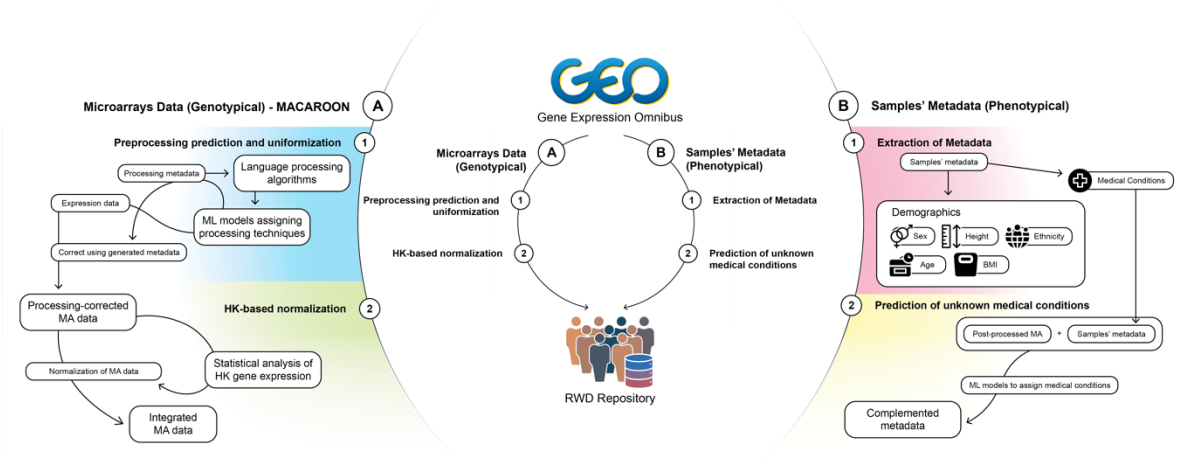

Figure 1. The overall pipeline of the methods applied in this work. A. Collection and processing of MA data. B. Collection and processing of RWD from GEO.

Global MA 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.

Text-mining of pre-processing procedures

Metadata accompanying the MA samples was downloaded. 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'll describe as pre-processing in the rest of this paper. A combination of terms of interest was used to select samples in datasets to train machine learning (ML) models. 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: application of Log2, Log10, Log

(unknown base), DE ratios, and Normalization. We then used these datasets for training ML models for inferring which pre-processing methods were applied to those cases in which metadata was not clear or was non-existent.

ML-based assignation of samples' pre-processing

The datasets mentioned above were used to train ML classifier models 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 combined it with a list of 5 low-variability HK genes (top five genes in **Table 3** in the Methods section). Best performing models are 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 (Eq. 1). In our results, models were capable of attaining satisfactory level of accuracy (cross-validation (CV); in the range [70%-98%]). However, to maintain maximum certainty in all predictions, we explored the accuracy of the top-three models in each quantile of the range of output score of the applied algorithm. We concluded 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 models, the model didn't display satisfactory prediction capabilities due to difficulties on clearly selecting samples on the negative class, which made us disregard these models in further steps. On the other hand, given the large number of missing values in the training datasets (especially concerning HK MA intensities), the various models were trained with a different sets 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 on 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 models 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**. After training, ML models were used to revert all samples to the original state by reverting the Log operation (see Methods section for detail). We decided to revert the Log transformation instead of applying it to all samples to take the mathematical nature of the Log operation into

account. Given that Log(x) doesn't exist if x is negative, to avoid approximation of imaginary numbers, values were restored to their original form.

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

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

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

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

*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

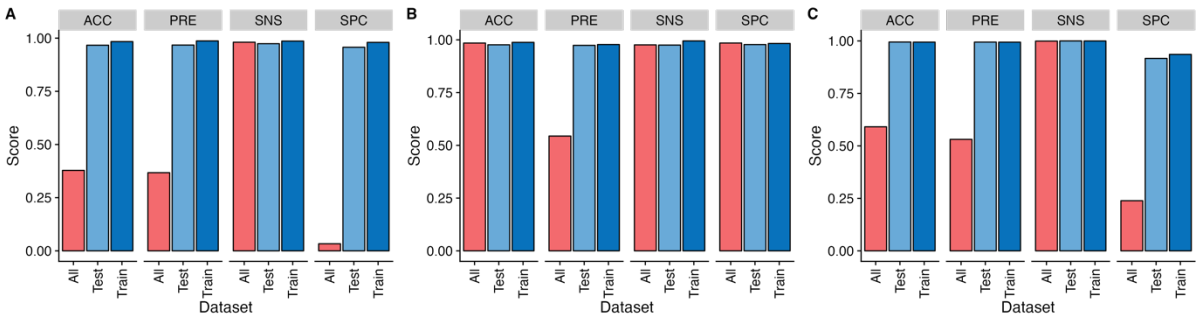

Figure 2. Global test metrics of the various ensemble models (A) Log2, (B) Log10, and (C) Ratio

HK-based regression for normalization

After the first stage of this method – aimed at cleaning and uniformizing original processing techniques, we designed a normalization method of MA data using regression algorithms from a 40 HK genes background dataset. This dataset is built using the median value of expression of the 40 genes 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, we were prevented from using such a strategy given that it doesn't allow for extrapolation. Given that HK genes are usually never under/overexpressed, extrapolation was a requirement in order to account for over-/under-expressed genes. For benchmarking, a test dataset containing MA data from prostate cancer samples and healthy controls (available in **Supplementary Table 3**) was extracted and their expression values were visually explored using 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[47], 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. Once again, we can see

that, with each step of Trans Babel, 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[21].

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

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

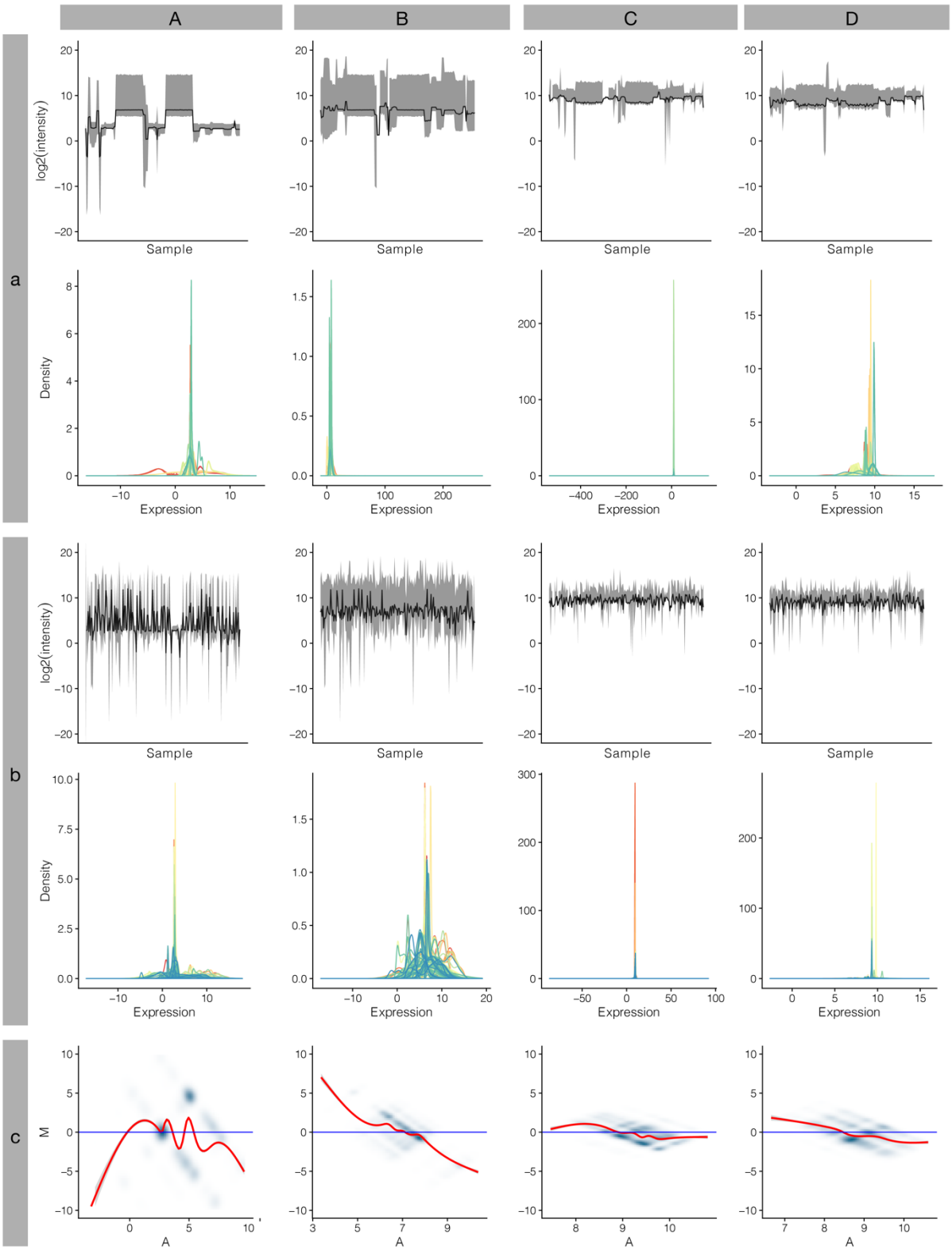

Figure 3. Log-expression data distribution across the various stages of MACAROON. A, B, C and D refer to the datasets produced along the process: A. Raw unprocessed data; B. After reverting original processing; C. After spline regression; and D. After polynomial regression. (a), (b), and (c) concern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

As means of benchmarking MACAROON, we compared its output against YuGene[48]: to our knowledge, the only available method that can also post-process MA (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[49] (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, we try to maintain the standard range of processed MA. In YuGene, values are scaled into the 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 observed later by 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. On the other hand, given that MA platforms have different coverages, 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.

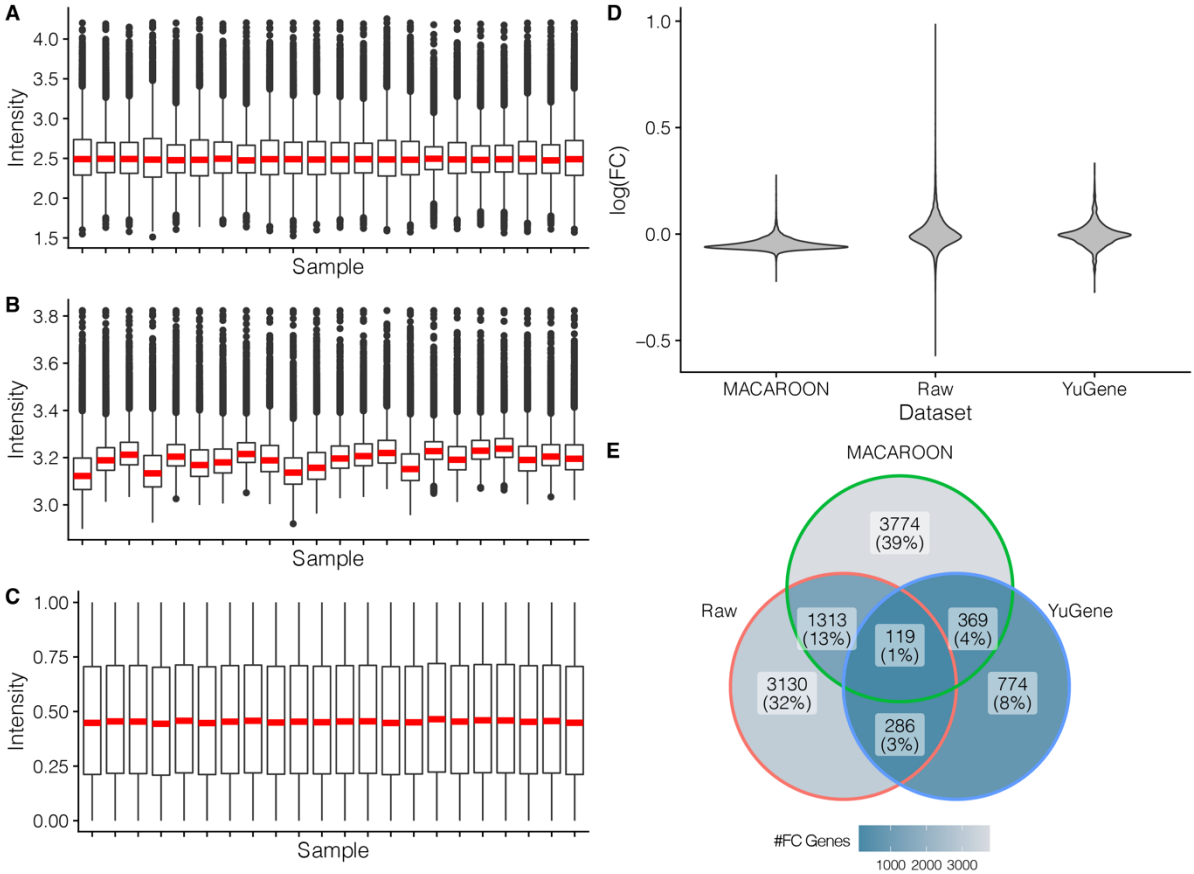

Figure 4. Direct comparison between methods. In the left column, we show the intensity per sample. A corresponds to the log2 raw data; B, using MACAROON; and C, using YuGene[48]. In the right column, we display a comparison of the results of a Differential Expression analysis. In D, we show a dispersion 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've used the same dataset used for the metrics presented in the previous sections in yet another DE Analysis experiment. We've used both spline and polynomial regressions. The overall results using each regression technique are displayed in **Figures 5A** and **B** via volcano plots, the full list of results is available as 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.[50], the intensity value detected by the scanner, is equivalent to the actual value of expression multiplied by two constants aimed at estimating the error at probe and sample levels (equation 7) – also a linear relationship. Significant differences in the DE results provided by the two methods rely 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.

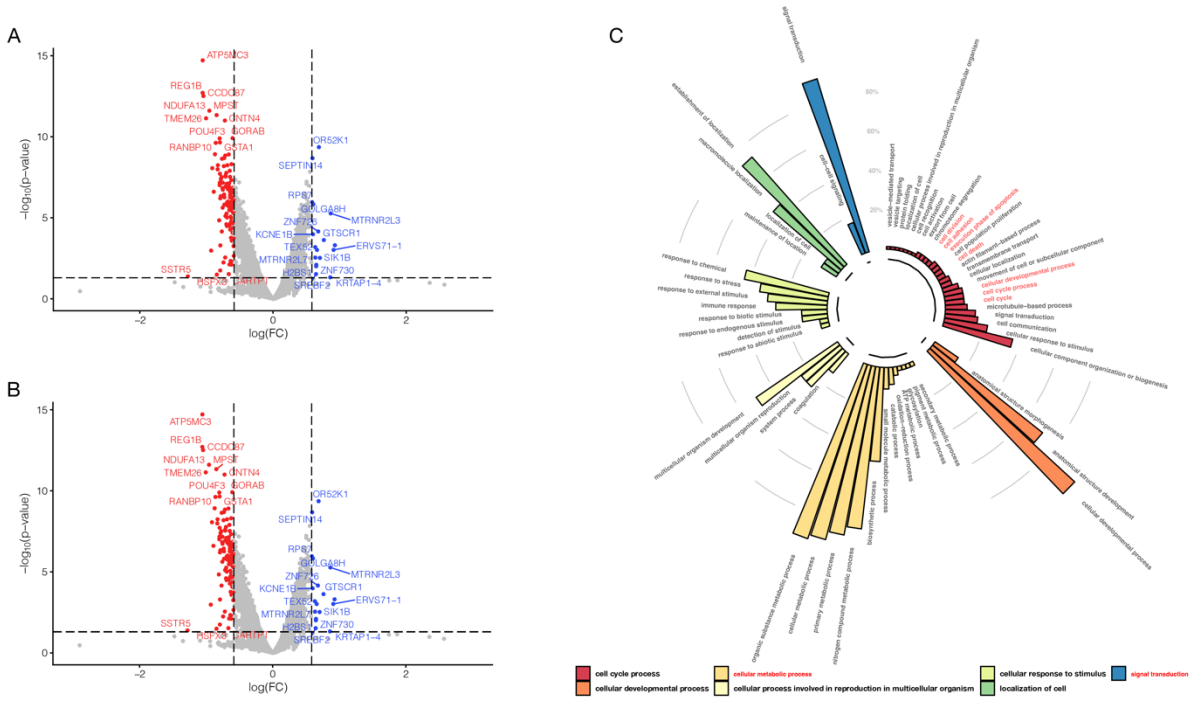

Figure 5. DE analysis results using spline and polynomial regressions. A, B. Volcano plots of the DE using both regression techniques. C. Enrichment analysis results using the polynomial dataset (where BPs found in other works in literature are highlighted in red). In A and B, the dots in red are under-expressed 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[51] Biological Processes (BP) using PANTHER[52]. Given, the theoretical linear behaviour of MA data discussed above, we've 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.*[53] reported differences in biological adhesion. Fan *et al.*[54] described changes in cell cycle control, cell death regulation, cell development, and focal adhesion. Sun *et al.*[55] and Verma *et al.*[56] 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 RWD repository

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[57,58], biological sex[59,60] or ethnical background [61,62]. Additionally, in the case of drug studies, information about body height is also crucial for the analysis/simulation of pharmacokinetics[63–65] and to better understand other unexpected effects (e.g. "BMI paradox"[66,67] – 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) and height, ethnicity, and ethnicity. Given the unstructured nature of metadata in GEO, not all 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**, we can find 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. A fact highlighted by the ethnological dispersion (**F6D**) of patients, supported by values of age, BMI, and height (**F6A**, **6B**, 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, we conclude that, 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.

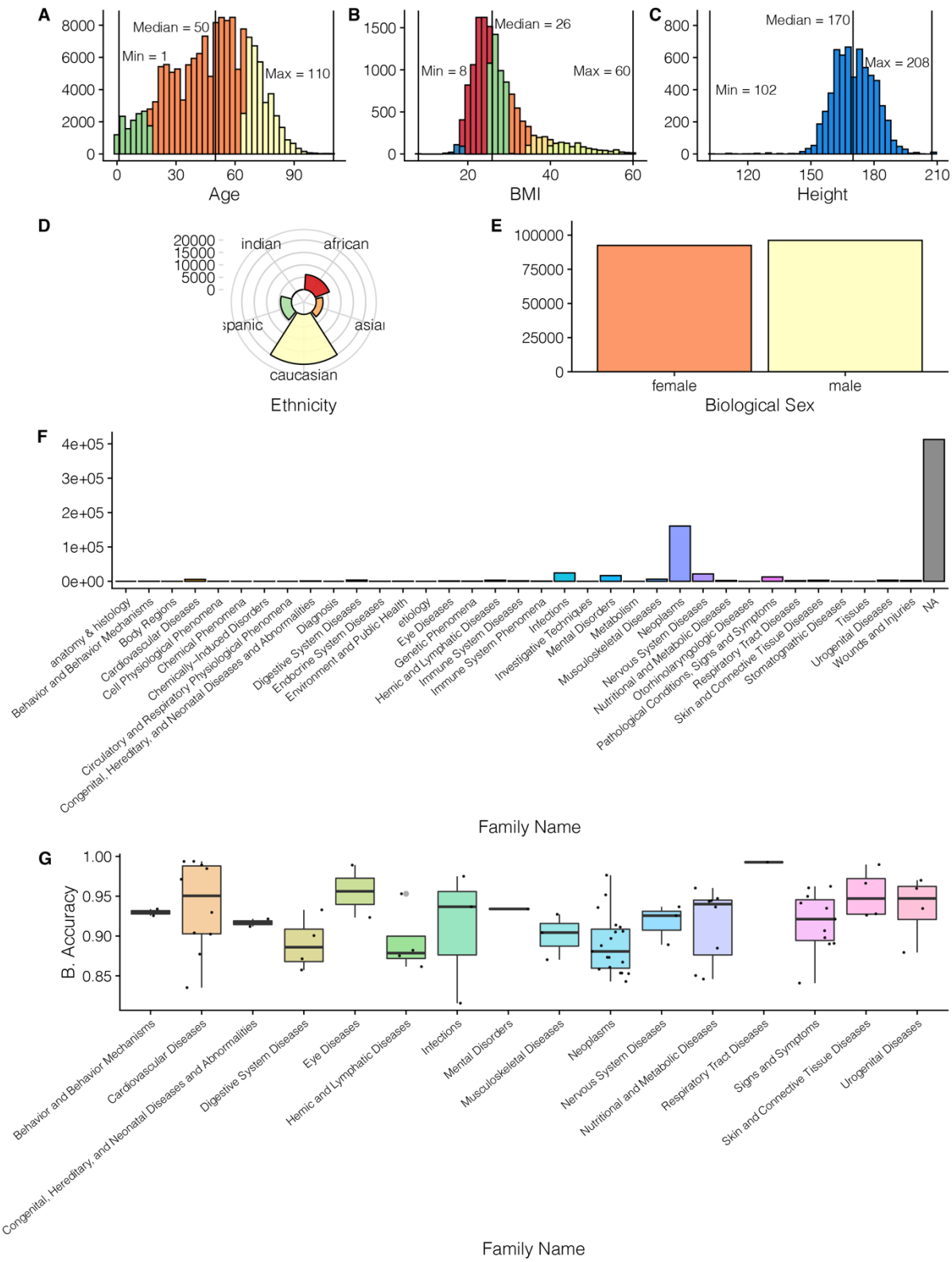

Figure 6. Demographic description of MA samples in GEO. A, Age distribution of samples (n=152,237). Colours represent various age groupings: green – childhood; orange – adulthood; yellow – elderhood. B, BMI distribution of samples (n=14,243). Different colours represent different gradings of BMI: blue – underweight; red – normal; orange – pre-obese; yellow – obese class I; green – 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 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[68], and OMiCC[69]) or by applying AI and NLP techniques (MetaSRA[70], and ALE[71]). 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[72]. In our dataset, we've 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.

A similar strategy was used to mine for information on medical conditions associated to the samples. The title of the study and the accompanying text was used to assign mentioned medical conditions and their status as healthy, control, or intervention case. Extracted pathological statuses were, then, mapped to the Biological Effectors database (BED)[73]. 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**). The BED 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. Linking these terms to BED permitted linking each medical condition, and, consequently, each sample in

GEO samples to molecular pathways of disease. The BED results of 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 errors in the database in comparison with other databases on the subject (e. g., DisGeNET[74] or KEGG[75]). **Figure 6F** describes the overall distribution of GEO patients across MeSH[76] conditions groups. As in the case of the demographical description, the distribution of diseases highlights the fact that 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.

To further complement medical condition annotation on the dataset, we resorted to ML classifiers. We trained a set of models by BED medical condition and used BED information on disease-related proteins to select the features for the ML models. The different models trained for each disease have a variable number of features, as means of counteracting the large number of missing probes (due to the different MA platform coverage – a subject discussed in previous sections of this work). A global depiction of the cross-validation balanced accuracy 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**. Across all available diseases, models reached at least 81.57% balanced accuracy (average: 91.47%; standard deviation: 0.044). However, we were unable to produce viable models for 59.02% of BED medical conditions. Of these, 63.19% were due to a lack of GEO samples from the corresponding BED medical condition. Regardless, the overall accuracy obtained through these models proves MACAROON's suitability for ML experiments on top of DE analysis. Additionally, these

models will also be used for filling the gaps in missing metadata, as well as studying possible comorbidities.

CONCLUSIONS

In this work, a global pipeline for the usage of GEO 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 profiles).

Given that, most data in GEO are available in its processed form, we start by presenting a method to ingrate MA data from multiple different sources onto a global MA dataset. Here, we aimed at being capable of using samples from previous experiments when raw (unprocessed) intensity values were not available, meaning when the authors only provide files processed and normalized. As previously explored, there are different processing techniques for the normalization of MA data, and, often, metadata supplementary to the datasets lacks the information needed to uniformize the datasets in Big Data experiments relying on data integration. To overcome these challenges, we design a processing protocol split into two steps: (1) metadata enrichment of samples using a natural-language processing algorithm; and (2) HK-based normalization using a cross-GEO background dataset. We've tested each intermediary step independently, and the results show great improvement as we've gone through each stage. The usage of MACAROON was, then, tested in a Prostate Cancer case study and could reproduce the biological conclusions found in other works in literature analysing the transcriptome of Prostate Cancer Samples. Such results support the usage of MACAROON in future transcriptomics experiments, providing a way of exploring new biological patterns in previously published data independently of previous processing it may have been subjected to, allowing the usage of bigger and more diverse cohorts – as required for the discovery of new, patient-centred biomarkers and pharmacological targets. MACAROON seems 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.

Secondly, 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, aiming at describing the patient population and allowing access to 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 certain distribution, according to the research purpose. Expectedly, the 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

Obtainment of the original GEO data

The transcriptomics dataset was constructed using MA data available in the GEO repository[30,77]. We've 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 files with data already processed by the researchers originally submitting them. Moreover, the text description of the data (included in the "Data table header descriptions" table of the sample page) was also kept and used in subsequent methods (this process is schematized in figure

S1A). This step was performed in Python 3.9.0 using the package Beautiful Soup 4[78]. We obtained a total of 1,866,292 containing processed MA data files. 211,384 samples did not contain any expression data.

As probe names change depending on the platform, probe ids were uniformized to reflect the corresponding gene. The soft file ("GPLXXX_family.soft.gz") corresponding to the platform mentioned on 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

All the samples downloaded in the previous step were statistically described, and information on the maximum value, minimum value, average, standard deviation, and variation was calculated. Additionally, the value corresponding to the expression of five HK genes was kept. HK genes are responsible for the basic functions of the cell[79] and are universally expressed in all tissue and cell types[80]. Given the known variation in the expression levels of some of these genes in some diseases or conditions[81], we increasingly ranked known genes according to the standard deviation of their expression across multiple tissues, as published by Eisenberg et al. (2013)[82]. Additionally, the ranked list was filtered to only include genes appearing at a minimum of 13 different previously available lists, as reported by Zhang et al. (2015)[83]. The final list of HK genes is available in **Table 3**. 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**.

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

Metadata description of the MA samples was processed using a text mining pipeline to build datasets that would allow the training of ML models to predict 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 recorded if the samples contained a series of "hot words": "calibrat" (this simplified word allows for the detection of various verbal forms of calibration), "intensity", "log", "log 10", "log 2", "log10", "log2", "mas", "normali" (this simplified word allows for the detection of American and British English, as well as various verbal forms normalization), "quantile", "ratio", "raw", and "rma". The dataset includes a sample if its description includes a "hot word", as portrayed in Table 3. 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 above were used to train ML classifiers. Given the reduced number of features and their separability (see figure S2), we used a simple set of models and selected the best-performing ones in the following steps: Decision tree, Elastic net, Elastic net with threshold optimization, Generalized linear model, Linear classifier, Multi-layer perceptron, Naïve Bayes Classifier, Quadratic classifier, and Support Vector Machines. Most models were written in MATLAB (version 9.2)[84] and its "Statistics and Machine Learning Toolbox"[85]. Models that were not available (Linear and Quadratic classifiers, and threshold optimization) were deployed using custom code. The hyperparameters of the applied models are described in **table 5**. Threshold optimization is done by calculation of the receiver operator curve (ROC). For storage optimization, all models attaining less than 0,70 validation accuracy were automatically deleted. The final models used in the classification phase comprise soft-voting ensemble models of 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*	-

Benchmarking of processing prediction models

Validation of ML models was performed by splitting the dataset (defined in sections above and according to **Table 4**) into training and test sets, thus presenting the models with samples it had never seen. 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

The raw score (model output before activation functions) from the models was recorded. This value was used to check the models' confidence in the prediction depending on the prediction score. Scores were separated into 20 intervals and the mean accuracy by interval was calculated. Only predictions falling into intervals with more than 95% accuracy are considered.

Uniformization of processing techniques

Taking the ML models described above, the MA values obtained in the first stages (already processed by the original researchers) were corrected to have them restored to their original form (before transformation). The transformations applied are described in equations 5 and 6.

Where log2 is detected,

$$x' = 2^x \qquad \text{(Eq. 5)}$$

And, where log10 is detected,

$$x' = 10^x \qquad \text{(Eq. 6)}$$

Cross-sample normalization method

After undoing logarithm transformations, we applied a normalization technique to uniformly transform all samples in the dataset. Given that different samples may have been acquired in very different settings (e.g., different MA scanners, processing software), and that we couldn't resort to control measurements in the samples, we decided to apply a normalization strategy that would level the various samples using a biologically intrinsic value in the samples. We've used the intensity values of HK genes to infer a sample-specific normalization curve that would be capable of estimating the expression of the remaining genes. The intensity value from all the downloaded samples of the 40 genes mentioned in Table 2 was captured. Given that none of the genes show a normal distribution of intensity, we've 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 and (2) polynomial regressors – 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 against other integration methods

We've tested MACAROON against a YuGene[48] – 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[49] – corresponding to the "Gene expression profiling of treated and untreated primary prostate cancer"). The three datasets were submitted to a DE analysis where differences between treated and untreated samples were compared. LogFC corresponds to the subtraction of the mean values of each cohort, p-values and false discovery rate (FDR) were calculated using a student's t-test. Differentially expressed genes were selected by excluding values with p-value or FDR >0.05.

Comparative case study

To evaluate the normalization technique, we've built a MA dataset containing 200 samples of Prostate Cancer from 63 GEO experiments (see **Supplementary Table 3** for further detail). A control dataset was built using the sample control samples from the GEO experiments. To account for possible errors in the ML-based stages of MACAROON, maximum values of expression in each sample were calculated, and samples corresponding to outlier maximum values were excluded. Differentially Expressed Genes (DEGs) were computed, and p-values were calculated using a two-sample T-test. The assessment of DEGs was performed via enrichment analysis for Gene Ontology BP[51] using PANTHER[52].

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 the metadata information provided alongside GEO samples. We mined information for medical conditions and demographical information. Additionally, bearing in mind the significant differences that exist between gene expression in tissues we felt the need to also assign samples' tissues. Information was mined from the text descriptions of GEO samples (specifically from the "GSEXXXXX-GPLXXXXX_series_matrix.txt files" files).

Demographical features

To define the dataset's samples demographically, we mined the provided metadata text for information on age, biological sex, ethnicity, BMI and height (acquired specifically from the fields "!Sample_characteristics_ch2" of the metadata file). Age was mined from the subfields "age" and "birth". The obtained raw text was then standardized to comprise the age of the subject in years using a custom MATLAB script. 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 mined from the fields "ethnicity", "race" and "population". Different vocabulary was standardized to "african", "asiatic", "caucasian", "hispanic" and "indian". Information on BMI was mined from the subfield "BMI". No transformations were performed in this class. 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 classified according to the medical condition provided in GEO series and the samples' titles and descriptions. These sentences were mined for a series of hot words. Medical conditions, corresponding to BED terms, were assigned to each sample by using a manually

curated synonym dictionary. Healthy and control samples were explicitly mined to account different experimental designs in GEO series. The different BED terms were mapped to MeSH terms[76] using MetaMap[86] – 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 fields "!Sample_characteristics_ch2" of the metadata file. To reduce redundancy, and to allow for hierarchical classification of body tissues, terms were mapped to the Brenda Tissue Ontology[87].

ML models to predict BED medical conditions

For each BED medical condition mined from the GEO samples, we tried to train a set of ML models aimed at deciphering if the remaining samples belong to that medical condition. Proteins associated with each medical condition in the BED were selected as features for the ML models. From the possible features, models were trained in a grid-search fashion, including feature selection (following the strategy applied to other ML models in this work). Models were trained with a maximum of 2,000 samples for each class, where the positive class contains GEO samples from which the BED medical condition was mined; and the negative class corresponds to a random set of samples where the medical condition was not mentioned. To account for variability in the number of samples, models were selected based on their cross-validation (k=10) balanced accuracy (equation 7), where only the models with balanced accuracy above 80% (p-value<0.01) were kept. To account for missing values in the datasets, for each medical condition, models had a different number of features (ranging from 2 to 5).

$$balanced\ accuracy = \frac{sensitivity + specificity}{2} \qquad \text{(Eq. 7)}$$

Abbreviations

ACC	Accuracy	HT	High-throughput
AI	Artificial Intelligence	MA	Microarrays
BED	Biological Effectors Database	MeSH	Medical Subject Headings
BMI	Body Mass Index	ML	Machine Learning
BP	Biological Process	NLP	Natural Language Processing
CUI	Condition ID	PRE	Precision
CV	Cross-validation	RCT	Randomized Clinical Trial
DE	Differential Expression	RNA-Seq	RNA Sequencing
DEG	Differentially Expressed Genes	ROC	Receiver-operator Curve
EMA	European Medicines Agency	RWD	Real-World Data
FC	Fold-change	SNS	Sensitivity
FDA	Food and Drug Administration	SPC	Specificity
FN	False Negatives	TN	True Negatives
FP	False Positives	TP	True Positives
GEO	Gene Expression Omnibus	UMLS	Unified Medical Language System
HK	House-Keeping		

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 doesn't 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 noninterventional 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

All data included in this work is publicly available at the GEO database – series applied in benchmarking and case studies are referenced in the methods section and 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 765912 (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, M. K. P., Xie, W., Ruder, L., Schulze, U. & Meier, C. AI in small-molecule drug discovery: a coming wave? *Nat Rev Drug Discov* **21**, 175–176 (2022).
2. Paul, D. *et al.* Artificial intelligence in drug discovery and development. *Drug Discov Today* **26**, 80–93 (2021).
3. Koscielny, G. *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res* **45**, D985–D994 (2017).
4. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence 2020 2:10* **2**, 573–584 (2020).
5. Rodriguez, S. *et al.* Machine learning identifies candidates for drug repurposing in Alzheimer's disease. *Nature Communications 2021 12:1* **12**, 1–13 (2021).
6. Rashid, M. B. M. A. *et al.* Optimizing drug combinations against multiple myeloma using a quadratic phenotypic optimization platform (QPOP). *Sci Transl Med* **10**, (2018).
7. Kolla, L., Gruber, F. K., Khalid, O., Hill, C. & Parikh, R. B. The case for AI-driven cancer clinical trials – The efficacy arm in silico. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1876**, 188572 (2021).
8. Collins, G. S. *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* **11**, e048008 (2021).
9. Park, S. H., Choi, J. & Byeon, J. S. Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence. *Korean J Radiol* **22**, 442–453 (2021).
10. Baumfeld Andre, E., Reynolds, R., Caubel, P., Azoulay, L. & Dreyer, N. A. Trial designs using real-world data: The changing landscape of the regulatory approval process. *Pharmacoepidemiol Drug Saf* **29**, 1201–1212 (2020).

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

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

13. 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.

14. Zura, R., Irwin, D. E., Mack, C. D., Aldridge, M. L. & Mackowiak, J. I. Real-World Evidence: A Primer. *J Orthop Trauma* **35**, (2021).

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

16. Shapiro, J. A. Revisiting the central dogma in the 21st century. *Ann N Y Acad Sci* **1178**, 6–28 (2009).

17. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science (1979)* **348**, 660–665 (2015).

18. Velculescu, V. E. *et al.* Characterization of the Yeast Transcriptome. *Cell* **88**, 243–251 (1997).

19. Bainbridge, M. N. *et al.* Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**, 1–11 (2006).

20. Fu, X. *et al.* Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10**, 1–9 (2009).

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

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

23. Minnier, J., Pennock, N. D., Guo, Q., Schedin, P. & Harrington, C. A. RNA-seq and expression arrays: Selection guidelines for genome-wide expression profiling. *Methods in Molecular Biology* **1783**, 7–33 (2018).

24. Thompson, J. A., Tan, J. & Greene, C. S. Cross-platform normalization of microarray and RNA-seq data for machine learning applications. *PeerJ* **2016**, e1621 (2016).

25. Belle, A. *et al.* Big data analytics in healthcare. *Biomed Res Int* **2015**, (2015).

26. Roski, J., Bo-Linn, G. W. & Andrews, T. A. Creating value in health care through big data: Opportunities and policy implications. *Health Aff* **33**, 1115–1122 (2014).

27. Chen, L. *et al.* Correlation between RNA-Seq and microarrays results using TCGA data. *Gene* **628**, 200–204 (2017).

28. Nookaew, I. *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 Res* **40**, 10084–10097 (2012).

29. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biol* **13**, e1002195 (2015).

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

31. Cheng, L., Lo, L. Y., Tang, N. L. S., Wang, D. & Leung, K. S. CrossNorm: a novel normalization strategy for microarray data in cancers. *Scientific Reports 2016 6:1* **6**, 1–11 (2016).

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

33. Carey, M., Ramírez, J. C., 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. *Stat Methods Med Res* **27**, 1930–1955 (2018).

34. Koeppen, K., Stanton, B. A. & Hampton, T. H. ScanGEO: parallel mining of high-throughput gene expression data. *Bioinformatics* **33**, 3500–3501 (2017).

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

36. Chen, G. *et al.* Restructured GEO: restructuring Gene Expression Omnibus metadata for genome dynamics analysis. *Database* **2019**, 145 (2019).

37. Khomtchouk, B. B., Dyomkin, V., Vand, K. A., Assimes, T. & Gozani, O. Biochat: a database for natural language processing of Gene Expression Omnibus data. *bioRxiv* 480020 (2018) doi:10.1101/480020.

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

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

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

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

42. Arisi, I. *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* 2735–2737 (2019) doi:10.1109/BIBM.2018.8621072.

43. Kato, S. *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* **11**, 1–9 (2020).

44. Fernandes, L. E. *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. *Clin Breast Cancer* **21**, e340–e361 (2021).

45. Wallach, J. D. *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 Netw Open* **4**, e2133667–e2133667 (2021).

46. Refaeilzadeh, P., Tang, L. & Liu, H. Cross-Validation. *Encyclopedia of Database Systems* 532–538 (2009) doi:10.1007/978-0-387-39940-9_565.

47. Vafaee, F., Dashti, H. & Alinejad-Rokny, H. Transcriptomic Data Normalization. in *Encyclopedia of Bioinformatics and Computational Biology* (eds. Ranganathan, S., Gribskov, M., Nakai, K. & Schönbach, C.) vol. 2 364–371 (Academic Press, 2019).

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

49. Sowalsky, A. G. *et al.* Neoadjuvant-Intensive Androgen Deprivation Therapy Selects for Prostate Tumor Foci with Diverse Subclonal Oncogenic Alterations. *Cancer Res* **78**, 4716–4730 (2018).

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

51. Carbon, S. *et al.* The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325–D334 (2021).

52. Mi, H. *et al.* PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res* **49**, D394–D403 (2021).

53. Yang, J., Chen, L., Kong, X., Huang, T. & Cai, Y. D. Analysis of Tumor Suppressor Genes Based on Gene Ontology and the KEGG Pathway. *PLoS One* **9**, e107202 (2014).

54. Fan, S. *et al.* Identification of the key genes and pathways in prostate cancer. *Oncol Lett* **16**, 6663–6669 (2018).

55. 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 Med Genet* **20**, 1–9 (2019).

56. Verma, S., Shukla, S., Pandey, M., MacLennan, G. T. & Gupta, S. Differentially expressed genes and molecular pathways in an autochthonous mouse prostate cancer model. *Front Genet* **10**, 235 (2019).

57. Montes, J. *et al.* Ambulatory function in spinal muscular atrophy: Age-related patterns of progression. *PLoS One* **13**, e0199657 (2018).

58. Przemska-Kosicka, A. *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. *Front Immunol* **8**, 591 (2018).

59. Basoglu, O. K. & Tasbakan, M. S. Gender differences in clinical and polysomnographic features of obstructive sleep apnea: a clinical study of 2827 patients. *Sleep and Breathing* **22**, 241–249 (2018).

60. Gabler, N. B. *et al.* Race and Sex Differences in Response to Endothelin Receptor Antagonists for Pulmonary Arterial Hypertension. *Chest* **141**, 20–26 (2012).

61. Morris, J. C. *et al.* Assessment of Racial Disparities in Biomarkers for Alzheimer Disease. *JAMA Neurol* **76**, 264–273 (2019).

62. Kopel, J. *et al.* Racial and Gender-Based Differences in COVID-19. *Front Public Health* **8**, 418 (2020).

63. Hebbes, C. P. & Thompson, J. P. Pharmacokinetics of anaesthetic drugs at extremes of body weight. *BJA Educ* **18**, 364–370 (2018).

64. Morrish, G. A., Pai, M. P. & Green, B. The effects of obesity on drug pharmacokinetics in humans. *http://dx.doi.org/10.1517/17425255.2011.570331* **7**, 697–706 (2011).

65. Smit, C., de Hoogd, S., Brüggemann, R. J. M. & Knibbe, C. A. J. Obesity and drug pharmacology: a review of the influence of obesity on pharmacokinetic and pharmacodynamic parameters. *https://doi.org/10.1080/17425255.2018.1440287* **14**, 275–285 (2018).

66. Tittl, L. *et al.* 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 (2018).

67. Donini, L. M., Pinto, A., Giusti, A. M., Lenzi, A. & Poggiogalle, E. Obesity or BMI Paradox? Beneath the Tip of the Iceberg. *Front Nutr* **7**, 53 (2020).

68. Hadley, D. *et al.* Precision annotation of digital samples in NCBI's gene expression omnibus. *Scientific Data 2017 4:1* **4**, 1–11 (2017).

69. Shah, N. *et al.* A crowdsourcing approach for reusing and meta-analyzing gene expression data. *Nature Biotechnology 2016 34:8* **34**, 803–806 (2016).

70. Bernstein, M. N., Doan, A. & Dewey, C. N. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics* **33**, 2914–2923 (2017).

71. Giles, C. B. *et al.* ALE: Automated label extraction from GEO metadata. *BMC Bioinformatics* **18**, 7–16 (2017).

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

73. Jorba, G. *et al.* In-silico simulated prototype-patients using TPMS technology to study a potential adverse effect of sacubitril and valsartan. *PLoS One* **15**, e0228926 (2020).

74. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* **48**, D845–D855 (2020).

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

76. National Center for Biotechnology Information & U.S. National Library of Medicine. MeSH (Medical Subject Headings). https://www.ncbi.nlm.nih.gov/mesh/.

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

78. Richardson, L. Beautiful soup documentation. Preprint at (2007).

79. Butte, A. J., Dzau, V. J. & Glueck, S. B. Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". *https://doi.org/10.1152/physiolgenomics.2001.7.2.95* **7**, 95–96 (2001).

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

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

82. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends in Genetics* **29**, 569–574 (2013).

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

84. The Math. MATLAB: r2017b. Preprint at (2017).

85. Mathworks. Statistics and Machine Learning Toolbox Documentation. Preprint at https://www.mathworks.com/help/stats/.

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

87. Gremse, M. *et al.* The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res* **39**, D507–D513 (2011).