Supplementary File 2: FORESEE Data Overview and Preparation

L. K. Turnhoff, A. H. Esfahani, M. Montazeri, N. Kusch and A. Schuppert

June 2018

1 Introduction

One of the major accommodations provided by FORESEE in facilitating a rapid drug prediction model development is its rich curated data sets. More than two gigabytes of cell line, xenograft and patient data sets are downloaded, curated and structured in objects, which are designed for more convenient model development. In this text we describe the FORESEE data objects ForeseeCell and ForeseePatient. We also provide information about data sets added to FORESEE in the structure of the aforementioned objects.

2 ForeseeCell Object

ForeseeCell objects are either cell line or xenograft data sets that are structured to be used as a TrainObject in the FORESEE pipeline. They usually contain variety of molecular data types for model training, and various response variables, which are different measures of drug responses.

2.1 ForeseeCell Structure

ForeseeCell is very similar to the list data type in R programming language; it is a data structure which includes different data types, and can be indexed using double brackets or dollar sign, for example, ForeseeCell$variable1 or ForeseeCell[["variable1"]] or ForeseeCell[[1]]. We divide components of ForeseeCell into two categories:

2.1.1 Fixed Components

These components are available in all instances of ForeseeCell. Here we list these fixed components:

InputTypes InputTypes is a data frame with two columns of 'Name' and 'Description', which provide the names of all components in the object that can be used as input data (in ForeseeTrain for example), and description for each input data.

ResponseTypes ResponseTypes is another two-column data frame with a 'Name' column, providing the names of all components in the object that are a measure of drug activity and can be used as response variable (called 'CellResponseType' in ForeseeTrain) and a 'Description' column for each response variable.

GeneExpression Althought technically part of the InputTypes, GeneExpression is present in all ForeseeCell object instances. It is a matrix, with genes in rows and samples (cell lines or xenograft model) in columns. Entrez IDs are saved in 'rownames' of the matrix and sample names in 'colnames'.

2.1.2 Potential Components

These components are not necessarily available in all instances of ForeseeCell. The potential components are as follows:

GeneExpressionRNAseq In data sets including both DNA array and RNA-seq measured gene expressions, DNA array values are saved in the 'GeneExpression' component and RNA-seq values are saved in 'GeneExpressionRNAseq'. Similar to GeneExpression, GeneExpressionRNAseq is a matrix, with genes in rows and samples (cell lines or xenograft model) in columns. Entrez IDs are saved in 'rownames' of the matrix and sample names in 'colnames'. Information about units used in GeneExpressionRNAseq can be found in the description of InputTypes. Availability of this potential component in a ForeseeCell data set can be checked via the InputTypes component.

ProteinExpression another matrix, with protein abundance values, with genes(corresponding to each protein) in rows and samples (cell lines or xenograft model) in columns. Entrez IDs are saved in 'rownames' of the matrix and sample names in 'colnames'. Information about technology and units used in ProteinExpression can be found in the description of InputTypes. Availability of this potential component in a ForeseeCell data set can be checked via the InputTypes component.

Methylation another matrix with genes in rows and samples (cell lines or xenograft model) in columns. Availability of this potential component in a ForeseeCell data set can be checked via the InputTypes component.

SNP6 Copy number variation measured by Affymetrix genome-wide human SNP Array 6.0 chip, arranged in a matrix similar to GeneExpression, genes in rows and samples in columns. Availability of this potential component in a ForeseeCell data set can be checked via the InputTypes component.

Mutation A binary matrix, with genes in rows and samples in columns, with one indicating existence of a mutation and zero indicating inexistence of a mutilation. Availability of this potential component in a ForeseeCell data set can be checked via the InputTypes component.

CNVGain A binary matrix, with genes in rows and samples in columns, where one indicates a severe increase in copy number variation and zero indicates normal copy number value. Availability of this potential component in a ForeseeCell data set can be checked via the InputTypes component.

CNVLoss A binary matrix, with genes in rows and samples in columns, where one indicates a severe reduction in copy number variation and zero indicates normal copy number value. Availability of this potential component in a ForeseeCell data set can be checked via the InputTypes component.

IC50,EC50,IG50,ActArea,BestResponse,... Response variables matrices, with samples in rows and drugs in columns. Availability of response components can be check via the ResponseTypes component, which also includes a description about the meaning behind each of the response measures.

DrugInfo A data frame with extra information about the included drugs in the dataset. Columns 'DRUG_NAME' and 'TARGET' from this component are used in FeatureSelector.ontology() and FeatureSelector.pathway() and are needed if the user wants to use any of the mentioned FeatureSelector methods, where the pipeline uses only the gene names for training the model that are contained in the ontology or pathway associated with the chosen drug.

CelllineInfo A data frame with extra information about cell lines in the dataset.

TissueInfo A data frame with tissue-related information about cell lines or xenograft models in the dataset. Column 'Site' of this component is used to extract relevant samples that user set by assigning a value to 'TrainingTissue' input in ForeseeTrain(). Hence, this component is needed if user wants to use a specific tissue for 'TrainingTissue'.

File	Name of the File	Last updated on	Link
log(IC50) and AUC values	v17.3_fitted_dose_response.xlsx	March 27th 2018	[1]
Screened compounds	Screened_Compounds.xlsx	March 27th 2018	[2]
Annotated list of Cell lines	Cell_Lines_Details.xlsx	July 4th 2016	[3]
RMA normalised gene expression from DNA array	sanger1018_brainarray_ensemblgene_rma.txt.gz	March 2nd 2017	[4]
Binary matrix of CNV, Mutation and Methylation	mmc4.xlsx	—	[5]

Table 1: GDSC Raw Files Overview

2.2 Available ForeseeCell Instances in the Package

2.2.1 GDSC

Genomics of Drug Sensitivity in Cancer, or GDSC for short, is one of the ForeseeCell datasets available in the FORESEE package. All files related to the GDSC dataset were downloaded on 25.4.2018 from https://www.cancerrxgene.org/downloads. Details of the downloaded files are listed in table 1.

Downloaded files are processed and prepared as the following components of the GDSC ForeseeCell object:

- **GeneExpression**: RMA normalized DNA array values were converted into an R matrix, with genes in rows and cell lines in columns. Column names that were originally COSMIC IDs were converted to cell line names, and row names were converted from Ensembl gene IDs to Entrez IDs using biomaRt [6, 7] (From here onward, any gene identifier conversion is done by biomaRt unless stated otherwise.).
- **CNVGain, CNVLoss, Mutation and Methylation**: Four different binary matrices, with genes in rows and cell lines in columns, all extracted from supplement Table S3B of the paper [8]. Gene names were converted from symbols to Entrez IDs.
- **LN_IC50, AUC ,RMSE ,Z_SCORE ,MAX_CONC_MICROMOLAR , MIN_CONC_MICROMOLAR**: Response data were all in one data frame. We rearranged drug responses into different matrices, with cell lines as rows and drugs as columns.
- **DrugInfo**: There are drugs that have two matching IDs: we added a '(2)' suffix to the drug name that already had another ID.
- **CelllineInfo and TissueInfo**: Directly imported from their corresponding files without any change.
- **InputTypes**: We listed all input data, and a short description for each, in one data frame.
- **ResponseTypes**: We listed all response data, and a short description for each, in one data frame.

For in depth look into the script that imports, processes and makes the GDSC object, check GDSCPreparer.R in the data-raw directory of the package.

2.2.2 CCLE

Broad Institute Cancer Cell Line Encyclopedia, or CCLE for short, is another big cell line dataset included as a ForeseeCell instance. All relevant files for the CCLE object were downloaded on May 2018 from https://portals.broadinstitute.org/ccle/data. You can check table 2 for detailed information about downloaded files.

Items inside CCLE object were built from the files in table 2, and are listed as follows (in depth built procedure can be found in CCLEPreparer.R):

- **GeneExpression**: A matrix, with gene Entrez IDs in rows and cell lines in columns, containing RMA-normalized gene expression profiles measured by DNA array.

File	Name of the File	Last updated on	Link
RMA normalised gene expression from DNA array	CCLE_Expression_Entrez_2012-10-18.res	October 18th 2012	[9]
RNAseq gene expression in RPKM	CCLE_DepMap_18Q1_RNAseq_RPKM_20180214.gct	February 14th 2018	[10]
Binary Calls for Copy Number and Mutation Data	CCLE_MUT_CNA_AMP_DEL_binary_Revealer.gct	February 29th 2016	[11]
Reverse Phase Protein Array data	ccle2maf_081117.txt	January 23rd 2018	[12]
Reverse Phase Protein Array antibody information	ccle2maf_081117-2.txt	January 24rd 2018	[13]
Cell Line Annotations	CCLE_sample_info_file_2012-10-18.txt	October 18th 2012	[14]
List of the 24 drugs profiled across 504 CCLE lines	CCLE_NP24.2009_profiling_2012.02.20.csv	April 17th 2012	[15]
Pharmacologic profiles for 24 anticancer drugs across 504 CCLE lines	CCLE_NP24.2009_Drug_data_2015.02.24.csv	February 24th 2015	[16]

Table 2: CCLE Raw Files Overview

- **GeneExpressionRNAseq**: A matrix, with genes identified by Entrez IDs in rows and cell lines in columns, containing gene expression profiles measured by RNA-seq, in Reads Per Kilobase of transcript per Million (RPKM).
- **Mutation, CNVGain and CNVLoss**: Three different binary matrices, with genes in rows (names converted to Entrez IDs) and cell lines in columns, pointing toward mutation, gaining copy number variation and losing copy number variation respectively.
- **ProteinExpression**: Another matrix, with genes in rows and cell lines in columns, containing values of protein abundance measured by Reverse Phase Protein Array. Gene corresponding to measured protein (Target genes of antibody) in each row is identified by its Entrez ID. Since there are duplications in measured genes, we used the mean value of the duplicated genes as the final value.
- **TissueInfo**: Directly imported as a data frame. The column 'Site Primary' was renamed to 'Site' so that TissueInfo would be compatible with SampleSelector() in case user assigned a 'TrainingTissue'.
- **DrugInfo**: Directly imported as a data frame. Columns 'Compound (code or generic name)' and 'Target(s)' were renamed to 'DRUG_NAME' and 'TARGET' respectively, for compatibility with FeatureSelector().
- **IC50, EC50, ActArea and Amax**: From CCLE response data, four matrices were build with cell lines in rows and drugs in columns. Descriptions about each measured value are available in ResponseTypes component.
- **ResponseTypes**: Data frame containing the names and a short description of all response components available in CCLE.
- **InputTypes**: Data frame containing the names and short description of all input data components available in CCLE.

2.2.3 DAEMEN

DAEMEN ForeseeCell contains the data used in [17]. We downloaded the data used for modeling, as provided by the paper with the link `https://www.synapse.org/#!Synapse:syn2179898`. An overview of all files downloaded for DAEMEN ForeseeCell are provided in table 3. Inside DAEMEN ForeseeCell we have:

- **GeneExpression**: Matrix of DNA array gene expressions, genes in rows and cell lines in columns, with Entrez IDs in rownames.
- **GeneExpressionRNAseq**: Matrix of counts based on RNA-seq technology. We transformed the seq values into logarithmic (base 2) scale for having semi-normal distributed values, which is necessary for linear modeling. As a prerequisite for transforming to log-scale, we replaced all values lower than 1 with 1.
- **Methylation and SNP6**: Directly imported into matrices from downloaded files.

File	Name of the File	Last updated on	Link
RMA normalised gene expression data from DNA array	Neve_AffyRMA_genelevel_maxvar_stringent.csv	August 30th 2013	[18]
Gene expression counts based on sequencing data	breastRNAseq_genelevel_stringent.txt	August 30th 2013	[19]
Methylation data	Methylation_stringent.csv	August 30th 2013	[20]
Methylation annotation data	Methylation_annotation_stringent.csv	August 30th 2013	[21]
SNP data	SNP6_genelevel_stringent_std0.7.csv	August 30th 2013	[22]
GI50 drug response	gb-2013-14-10-r110-S1.xlsx	—	[23]

Table 3: DAEMEN Raw Files Overview

- **GI50**: Directly imported as a matrix with cell lines in rows and drugs in columns. A description about GI50 can be found in ResponseTypes component.
- **TissueInfo**: Directly imported as a data frame. The column 'Transcriptional subtype' was renamed to 'Site' to make TissueInfo compatible with SampleSelector().
- **ResponseTypes**: Data frame containing the names (only GI50 in this data set) and a short description of all response components available in DAEMEN.
- **InputTypes**: Data frame containing the names and a short description of all input data components available in DAEMEN.

You can check DAEMENPreparer.R for more details on how the DAEMEN data set was built.

2.2.4 GAO

GAO is one of the two xenograft data sets included in FORESEE. The data is downloaded as supplement files of [24], which are freely available and can be downloaded via [25]. GAO ForeseeCell includes:

- **GeneExpression**: A matrix of RNA-seq values in FPKM (Fragments Per Kilobase of transcript per Million). The matrix contains genes as rows and samples as columns, with Entrez IDs in rownames. FPKM values were transformed into logarithmic scale (base 2) for having semi-normal distributed values, which is necessary for linear modeling.
- **SNP6**: Copy number data measured by SNP array (Affymetrix genome-wide human SNP Array 6.0 chip).
- **Mutation, CNVGain and CNVLoss**: Binary matrices, pointing toward mutations, gaining copy number variations and losing copy number variations respectively.
- **TissueInfo**: Directly imported from 'PCT raw data' sheet of the excel file as a data frame. The column 'Tumor Type' was renamed to 'Site' to make TissueInfo compatible with SampleSelector() in case user assigned a 'TrainingTissue'.
- **DrugInfo**: Directly imported from 'PCT curve metrics' sheet of the excel file as a data frame. Columns 'Treatment' and 'Treatment Target' were renamed to 'DRUG_NAME' and 'TARGET' respectively, for compatibility with FeatureSelector().
- **BestResponse, BestResponseCombo, TimeToDouble, ...**: This data set includes 14 different response matrices, all of which have samples in rows and drugs in columns. List and description of these response matrices can be found in ResponseTypes component.
- **ResponseTypes**: Data frame containing the names and a short description of all response components available in GAO.

- **InputTypes**: Data frame containing the names and a short description of all input data components available in GAO.

More details about how GAO data set was built can be found in GaoPreparer.R.

2.2.5 WITKIEWICZ

WITKIEWICZ is the other xenograft data set included in FORESEE. This data set is from a study by Witkiewcz et al. [26], studying Pancreatic ductal adenocarcinoma (PDAC) drug response. Data used in building the WITKIEWICZ ForeseeCell are two excel files, which are included as supplements in the original paper (accessible with the links [27, 28]) and the GEO data set GSE84023 (accessible via [29]), which includes the RNA-seq gene expression relevant to the paper. WITKIEWICZ data set contains:

- **GeneExpression**: Matrix of gene expressions measured by RNA-seq. We used the already processed data available on GEO. Based on GSE84023 page on GEO, this is the processing pipeline they used: "Illumina Casava1.7 software used for basecalling. Sequenced reads were trimmed for adaptor sequence, mapped to hg19 genome using bowtie TopHat. Counts per gene was obtained using HTseq counts and normalized using edgeR package in R. Genome_build: hg19. files_format_and_content: tab-delimited text file include matrix of normalized log counts per million for each sample."

  We averaged over all samples from the same patient.

- **AUC and AUCCombo**: Response data were imported from supplement excel files, and then formatted as a matrix with samples in rows and drugs in columns. More information about these two response matrices can be found in ResponseTypes component.

  In AUC matrix, we averaged over all samples from the same patient.

- **DrugInfo**: A data frame with only one column, which includes drug names of this study, and their alternative names (e.g. commercial name) in parentheses.

- **ResponseTypes**: Data frame containing the names and a short description of all response components available in WITKIEWICZ.

- **InputTypes**: Data frame containing the names (in this case only 'GeneExpression') and a short description of all input data components available in WITKIEWICZ.

You can check WITKIEWICZPreparer.R for more detail about how exactly this data set was processed.

3 ForeseePatient Object

ForeseePatient objects are, as the name suggests, patient data sets that are structured to be used as a TestObject in the FORESEE pipeline. In comparison to ForeseeCell, ForeseePatient data sets tend to have less components; they usually have only one input data type variable and one response variable.

3.1 ForeseePatient Structure

ForeseePatient, similar to ForeseeCell, is a data structure having components of different data types, that can be indexed using double brackets or dollar sign (for example ForeseePatient$variable1 or ForeseePatient[["variable1"]] or ForeseePatient[[1]]), similar to a list data type in R programming language. Analogous to ForeseeCell, we can divide the components of ForeseePatient into two categories:

3.1.1 Fixed Components

These components are available in all instances of ForeseePatient:

GeneExpression is a matrix, with genes in rows and patients in columns. Entrez IDs are saved in 'rownames' of the matrix and patient identifiers in 'colnames'.

Annotation is a logical or numeric vector indicating the patient response to a drug. Extra information is provided in names(Annotation), e.g. when Annotation is a logical vector, names(Annotation) provides information about what True and False in Annotation mean in terms of patient response.

Unlike ForeseeCell, ForeseePatient data sets do not have 'ResponseTypes' or 'InputTypes' components, since in all available ForeseePatient instances in the package, there is only one input type data ('GeneExpression') and only one response type data ('Annotation').

3.1.2 Potential Components

This component is not necessarily available in all instances of ForeseePatient:

ExtraAnnotation is a data frame including all annotations that was contained in the original patient data set. This component is not used in the FORESEE pipeline, but is included for the user (e.g. to divide a patient data set into sub groups based on ExtraAnnotation for better modeling).

3.2 Available ForeseePatient Instances in the Package

We included and preprocessed all patient data in the [30] in our package. Since the preprocessing and constructing pipeline for all ForeseePatient data sets are almost the same, the following explanation applies to all ForeseePatients:

- **GeneExpression**: Raw CEL files were downloaded from GEO, and normalized using RMA from affy package [31]. Except in case of GSE9782, which didn't provide raw CEL files on GEO, and for that we downloaded their already MAS5-normalized data. In contrast to [30] we did not use MAS5 normalized version as they were, but for better modeling purposes, we transformed the data to a logarithmic scale (base 2).

  Data sets that contained more than one drug types were split into smaller ForseePatient data sets, for example, in GSE33072 includes two groups of patients treated with erlotinib or sorafenib, therefore we divided it into two GSE33072_erlotinib and GSE33072_sorafenib ForeseePatient data sets.

- **Annotation**: extracted from provided annotations from GEO, similar to [30].

An overview of all ForeseePatient data sets are provided in table 4.

ForeseePatient data set Name	Treatment	Number of Patients	Link
GSE6434	Docetaxel	24	[32]
EGEOD18864 (identical to GSE18864 on GEO)	Cisplatin	24 (only patients, not reference tumors)	[33]
GSE33072_erlotinib	Erlotinib	25	[34]
GSE33072_sorafenib	Sorafenib	39	[34]
GSE9782_GPL96_bortezomib	Bortezomib	169	[35]
GSE9782_GPL96_dexamethasone	Dexamethasone	70	[35]
GSE9782_GPL97_bortezomib	Bortezomib	169	[35]
GSE9782_GPL97_dexamethasone	Dexamethasone	70	[35]

Table 4: Overview of Available ForeseePatient Instances

References

[1] ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/v17.3_fitted_dose_response.xlsx.

[2] ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/Screened_Compounds.xlsx.

[3] ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/Cell_Lines_Details.xlsx.

[4] ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/sanger1018_brainarray_ensemblgene_rma.txt.gz.

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4967469/bin/mmc4.xlsx.

[6] Steffen Durinck, Paul T. Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, 4:1184–1191, 2009.

[7] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21:3439–3440, 2005.

[8] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

[9] https://data.broadinstitute.org/ccle_legacy_data/mRNA_expression/CCLE_Expression_Entrez_2012-10-18.res.

[10] https://data.broadinstitute.org/ccle/CCLE_DepMap_18Q1_RNAseq_RPKM_20180214.gct.

[11] https://data.broadinstitute.org/ccle_legacy_data/binary_calls_for_copy_number_and_mutation_data/CCLE_MUT_CNA_AMP_DEL_binary_Revealer.gct.

[12] https://data.broadinstitute.org/ccle/CCLE_RPPA_20180123.csv.

[13] https://data.broadinstitute.org/ccle/CCLE_RPPA_Ab_info_20180123.csv.

[14] https://data.broadinstitute.org/ccle_legacy_data/cell_line_annotations/CCLE_sample_info_file_2012-10-18.txt.

[15] https://data.broadinstitute.org/ccle_legacy_data/pharmacological_profiling/CCLE_NP24.2009_profiling_2012.02.20.csv.

[16] https://data.broadinstitute.org/ccle_legacy_data/pharmacological_profiling/CCLE_NP24.2009_Drug_data_2015.02.24.csv.

[17] Anneleen Daemen, Obi L Griffith, Laura M Heiser, Nicholas J Wang, Oana M Enache, Zachary Sanborn, Francois Pepin, Steffen Durinck, James E Korkola, Malachi Griffith, et al. Modeling precision treatment of breast cancer. *Genome biology*, 14(10):R110, 2013.

[18] https://www.synapse.org/Portal/filehandleassociation?associatedObjectId=syn2184894&associatedObjectType=FileEntity&fileHandleId=150628.

[19] https://www.synapse.org/Portal/filehandleassociation?associatedObjectId=syn2184895&associatedObjectType=FileEntity&fileHandleId=150630.

[20] https://www.synapse.org/Portal/filehandleassociation?associatedObjectId=syn2184893&associatedObjectType=FileEntity&fileHandleId=150626.

[21] https://www.synapse.org/Portal/filehandleassociation?associatedObjectId=syn2184892&associatedObjectType=FileEntity&fileHandleId=150624.

[22] https://www.synapse.org/Portal/filehandleassociation?associatedObjectId=syn2184911&associatedObjectType=FileEntity&fileHandleId=150660.

[23] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3937590/bin/gb-2013-14-10-r110-S1.xlsx.

[24] Hui Gao, Joshua M Korn, Stéphane Ferretti, John E Monahan, Youzhen Wang, Mallika Singh, Chao Zhang, Christian Schnell, Guizhi Yang, Yun Zhang, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature medicine*, 21(11):1318, 2015.

[25] https://media.nature.com/original/nature-assets/nm/journal/v21/n11/extref/nm.3954-S2.xlsx.

[26] Agnieszka K Witkiewicz, Uthra Balaji, Cody Eslinger, Elizabeth McMillan, William Conway, Bruce Posner, Gordon B Mills, Eileen M O'Reilly, and Erik S Knudsen. Integrated patient-derived models delineate individualized therapeutic vulnerabilities of pancreatic cancer. *Cell reports*, 16(7):2017–2031, 2016.

[27] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5287055/bin/NIHMS803574-supplement-2.xlsx.

[28] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5287055/bin/NIHMS803574-supplement-3.xlsx.

[29] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84023.

[30] Paul Geeleher, Nancy J Cox, and R Stephanie Huang. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome biology*, 15(3):R47, 2014.

[31] Laurent Gautier, Leslie Cope, Benjamin M. Bolstad, and Rafael A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[32] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6434.

[33] https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-18864/.

[34] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33072.

[35] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9782.