

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Signature Search Method Development and Application in Drug Discovery

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics and Bioinformatics

by

Yuzhu Duan

March 2022

Dissertation Committee:

Dr. Thomas Girke, Chairperson

Dr. Wenxiu Ma

Dr. Giulia Palermo

Copyright by
Yuzhu Duan
2022

The Dissertation of Yuzhu Duan is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

First, I want to express my most sincere gratitude to my advisor Dr. Thomas Girke, who provided valuable guidance and help for my research and publications. I also want to thank the members of my qualifying exam and dissertation committees for their invaluable support, including Dr. Wenxiu Ma, Dr. Zhenyu Jia, Dr. Giulia Palermo, Dr. Ted Karginov, Dr. Caroline Roper, Dr. Shizhong Xu, and Dr. Jiayu Liao. As chair of my qualifying exam, Dr. Zhenyu Jia was very supportive and guided me through the exam process.

I also want to thank my lab members Dr. Brendan Gongol, Dr. Daniela Cassol, Jianhai Zhang and Le Zhang for their kindness and continuous support.

I also acknowledge the Bioconductor core team and community for providing valuable input for developing the *signatureSearch* and *signatureSearchData* packages, as well as the funding provided by the National Institute on Aging at the National Institutes of Health, and by the National Science Foundation.

Finally, I want to express my sincere love and thanks to my family. They gave me strength and encouragement whenever I needed it. They are the hope of my life.

The text of this dissertation, in part, is a reprint of the material as it appears in “signatureSearch: environment for gene expression signature searching and functional interpretation”, *Nucleic Acids Research*, December 2020. The co-author Thomas Girke listed in that publication directed and supervised the research which forms the basis for this dissertation.

ABSTRACT OF THE DISSERTATION

Signature Search Method Development and Application in Drug Discovery

by

Yuzhu Duan

Doctor of Philosophy, Graduate Program in Genetics, Genomics and Bioinformatics
University of California, Riverside, March 2022
Dr. Thomas Girke, Chairperson

This dissertation is about the development of gene expression signature (GES) search methods and their application in drug discovery, specifically for promoting healthy aging. GES searching is a powerful technology facilitating the identification of drugs for treating diseases and drug repurposing. This is achieved by identifying drugs in GES databases inducing signatures similar to query GESs obtained from diseased samples or drug treatments. The new connections are useful for developing pharmacological interventions. This dissertation is divided into the following three components. First, I developed the *signatureSearch* R/Bioconductor package that integrates existing and novel methods for GES searching and functional enrichment analysis (FEA). Subsequently, I tested the performance of different GES search methods. They represent the first systematic performance tests of these methods in the field. Second, I applied *signatureSearch* to the human healthy aging field to reveal insights into longevity associated (LA) drugs and their targets by searching the Integrated Network-based Cellular Signatures (LINCS) database. For this, I assessed the performance of LINCS drugs, inducing GESs representative for their

mechanism of action (MOA), by computing for each MOA a recall score based on the GES similarity of the corresponding drugs. The obtained recall scores were used to prioritize LA drugs in the downstream discovery. LA MOA categories along with the corresponding drugs were identified by querying LINCS with GESs of drugs present in LINCS and DrugAge, and scoring the enrichment of each MOA. The corresponding LA pathways were identified via global mapping of targets of LA drugs and MOA categories. Next, I searched LINCS with the GESs from 11 well-studied LA drugs and one longevity phenotype. To identify LA pathways, the targets of the newly identified LA drugs were used for FEA. The results from the three steps were integrated and then interrogated with a combinatorial approach to select the most reliable set of LA drugs and targets. Collectively, this study identified a list of drugs, targets and pathways useful for pharmacological lifespan extension strategies. Third, I developed several data packages to incorporate in *signatureSearch* detailed annotations of drugs and targets from different community databases.

Table of Contents

List of Figures	ix
List of Tables	xv
1 Introduction	1
1.1 Overview	1
1.2 Need to Develop GESS Software Environment	6
1.3 Discovery of Healthy Aging Drugs	7
1.4 Integration of Annotations from Other Sources	10
2 signatureSearch Tool	13
2.1 Abstract	13
2.2 Materials and Methods	14
2.2.1 Implementation	14
2.2.2 Data Types of Queries and Databases	14
2.2.3 Reference Databases	15
2.2.4 Pre-processing and cutoffs for queries and databases	16
2.2.5 Compatibility Among Data Types	19
2.2.6 Overview of Analysis Workflow	20
2.2.7 Analysis Methods	21
2.2.8 Software Design	35
2.3 Results	38
2.3.1 Performance Comparisons of GESS Methods	38
2.3.2 Use Case	48
2.4 Discussion	54
2.5 Conclusion	56
2.6 Availability of Software and Data	57
3 Application to Human Longevity	58
3.1 Abstract	58
3.2 Results	59
3.2.1 Recall Performance of MOA Categories	60

3.2.2	MOAs Connected with Longevity	66
3.2.3	LAD Targets within Global Pathway Map	69
3.2.4	Optimization of GESS/FEA Workflow	73
3.2.5	GESS and FEA Results for Eleven LAD Signatures	73
3.2.6	GESS/FEA Results for Longevity-based Phenotype Signature	86
3.3	Discussion	96
3.4	Materials and Methods	105
3.4.1	Gene Expression Signature Searching	105
3.4.2	Functional Enrichment Analysis	106
3.4.3	Eleven LADs Selection	106
3.4.4	MOA Annotation and Size Cutoff	107
3.4.5	Recall Performance of MOA Categories	108
3.4.6	MOAs Connected with Longevity	111
3.4.7	Drug-Target Pathway Map Plotting	112
3.4.8	Permutation Tests	112
3.4.9	Drug Prioritization Strategy	113
3.4.10	Voting Strategy on FEA Results	116
3.5	Supplementary Material	118
3.5.1	Supplementary Methods	118
3.5.2	Supplementary Sections	122
3.5.3	Supplementary Figures	126
3.5.4	Supplementary Tables	126
3.6	Data Availability	126
4	Annotation Packages	139
4.1	Abstract	139
4.2	Materials and Methods	140
4.2.1	Implementation	140
4.2.2	Software Design	141
4.3	Results	142
4.3.1	drugbankR Package	142
4.3.2	Shiny Web Application for Gene Target Annotation	144
4.3.3	customCMPdb package	147
4.4	Discussion	154
5	Conclusion	157
	Bibliography	160

List of Figures

2.1	Overview of GESS and FEA workflow. GES queries are used to search a drug-based GES reference database for drugs inducing GESs similar to the query. To interpret the results mechanistically, the GESS results are subjected to functional enrichment analysis (FEA) including drug and target set enrichment analyses (DSEA, TSEA). Both identify enriched functional categories (GO terms and/or KEGG pathways) in the GESS results. Subsequently, drug-target networks (DTNs) are reconstructed for visualization and interpretation.	22
2.2	Design of <i>signatureSearch</i> package. GES reference databases are constructed from expression profile collections (RNA-Seq, Affymetrix chip or other technologies) and stored as HDF5 files. To perform GESSs, all query parameters are defined in a <i>qSig</i> search object where users can choose among over five search algorithms. The results are stored in a <i>gessResult</i> object that can be functionally annotated with different TSEA and DSEA methods. The enrichment results are organized in an <i>feaResult</i> object that can be used for drug-target network analysis and visualization.	36
2.3	Performance testing strategy of GESS methods. (A) The GESs of the drugs in each MOA and SSC category were searched against the LINCS database with each of the six GESS methods. The results were sorted by the corresponding similarity scores, here indicated by boxes with color gradient. GESs from the same and different MOA/SSC categories (CAT) as the query were indicated in a binary vector with ones and zeros (next to boxes), respectively. After joining the binary vectors for each category group and re-sorting them by the corresponding scores, cumulative TPRs and FPRs were plotted in form of ROCs. This was done on the global level (B) and the CAT level (C) for the MOA and SSC classifications separately. (D) The distributions of AUC/pAUC values from each CAT-level are depicted by violin plots with mean values and standard deviation (STDEV) bars given in the middle. In addition, the global AUC/pAUC values are indicated by triangles. (E) The statistical significance of the observed differences among the global AUC/pAUC values of the six GESS methods was assessed by a bootstrap test described in the text.	43

2.4	Recall performance of GESS methods on MOA and SSC categories. (A) The distributions of the ROC performance results of the 69 MOA categories are plotted in form of violin plots for each of the six GESS methods. The corresponding mean values, standard deviation bars and global AUCs are indicated within each violin by dots, vertical lines and triangles, respectively. The GESS methods are ordered by increasing global AUC values. (B) The corresponding distributions of pAUC values are given for FPRs of 1%, 5% and 10%. In this composite plot, the GESS methods are ordered by the mean of the ranks of their global pAUC values. (C)-(D) The GESS performance results of the 139 SSC categories are plotted the same way as the corresponding MOA results. (E) The performance results under (A)-(D) are summarized in form of stacked bar plots where the sum of the ranks is used to order the GESS methods from left to right by increasing performance. Each bar is composed of the ranking of the global AUCs and the mean ranking of the corresponding pAUCs for both MOA and SSC categories.	45
2.5	Structure-based hierarchical clustering dendrogram for drugs listed in Table 2.6. Experimental drugs lacking structure information are not included. . . .	50
2.6	Drug-target network module of Histone Deacetylase Activity (H3-K14 specific; GO MF ID: GO:0031078). Drugs and targets are depicted as boxes and circles, respectively. The color of the circles indicates the number of connections.	53
3.1	Strategy for identifying novel LADs, LAGs and LAPs. (A) The recall performance of MOA categories with at least 5 annotated drugs were estimated by searching the corresponding GESs against LINCS. The same MOAs were also ranked by their connectivity to longevity by using the GESs of known LADs as queries (B). The corresponding LADs included those present in both DrugAge and LINCS, as well as a custom set of 11 core LADs. Longevity associated target pathways were identified by enrichment analysis using Reactome where the target proteins of the above LAD sets were used as test sets (C). (D-F) An optimized GESS/FEA workflow was applied to three different GES sets from: (i) three well-characterized drugs as proof of concept experiment; (ii) 11 core LADs; and (iii) a longevity phenotype. The voting strategy was applied on 11 LADs GESs to get prioritized LADs candidates. The candidates can be flagged and associated with MOAs in the above recall performance tests.	61

3.2	MOA recall rates and longevity association plots. (A) Top 20 MOAs ranked by recall performance. (B) Top 20 MOAs related to longevity ranked by median absolute CORct scores from GESSs using LAD87. (C) Top 20 MOAs related to longevity from GESSs using LAD11. Only MOAs with a minimum of 5 drugs are included for a total of 138 MOAs. Bold names indicate MOAs present in at least 2 panels. The complete MOA ranking results are available in Table S2, S3 and S4. ColMedianCORct values indicate the distribution of median absolute CORct for MOA drugs. Recall Rate are expressed as rank percentiles, N drugs represents the number of drugs in each MOA, and MedianCORct values indicate the median absolute CORct from LADs GES queries.	65
3.3	Global pathway mapping of drug targets. (A) Fireworks plot of the 28 highest level pathways in Reactome. Blue branches indicate enrichment p-values ≤ 0.01 for targets of the 11 LAD set. (B) Heatmaps of enrichment results for 28 high level pathways (rows) and 3 target sets (columns) are given for adjusted p-values of hypergeometric distribution test and normalized pathway mappings on left and right, respectively. Adjacent bar plots give the number of targets for each query set and the total number of genes in each pathway.	72
3.4	GESS result summaries across LAD11 queries. A: NCS score distributions in the GESS results from LAD11 queries and one random GES query as negative control for NCS scores less than -1.00 (left panel) and greater than 1.00 (right panel) after setting count cutoff as 100 to better show the NCS distributions at left and right extremes. The color key shows the P-values of the WTCS score for the entries in the GESS results. B: GESS result rankings of drugs in the same MOA as the query LAD. Numerical values in the x-axis labels indicate the number of drugs sharing the MOA. Black dots represent drugs sharing the same MOA as query LAD and Grey bars indicate the total number of drugs in GESS results. C: hierarchical clustering of LAD11 by their GESS results ranking similarity. The color key indicates the Spearman correlation coefficient of the GESS result rankings from NCS scores after filtering of zeros. D: Clustering of in vitro and in vivo samples from sirolimus, acarbose and estradiol queries by their GESS results ranking similarity. The color key is the same as Figure C. E and F: Top 50 PDs from voteNCSunique method with stratification on LADs queries of sirolimus and acarbose, respectively. The columns are query samples that are clustered by Euclidian distance of NCSunique scores. The compound annotations plotted in the right bars include layer information (Layer), whether in DrugAge database (isLAD), whether the compound share at least one MOAs with MOAs of LAD11 (MOA Match), the number of compound targeted Reactome pathways shared with target Reactome pathways of LAD11 (N Pathway), whether the compound is therapeutic (Therapeutic), Max phase study by FDA (Max Phase). The complete tables containing all compounds rankings from voting strategy with scores, layer and annotation information corresponding to the heatmaps are stored in Synapse (syn27074560).	79

3.5	Combined PDs from LAD11 queries with at least 2 LADs support. The top 50 PDs from each individual LAD of LAD11 are combined into one drug list ranked by number of LADs queries that have the drug in their top 50 PD list (Nsupport) and the drugs are filtered with at least 2 Nsupport. It results in 91 drugs in this list. The corresponding table with no Nsupport cutoff is at Table S8.	80
3.6	FEA results summary. A: ranking positions of Reactome pathways meeting a 0.05 adjusted p-value cutoff from targets of the query LADs in their FEA results. Black dots represent pathways in FEA results matching the direct pathways from the query LAD targets. Grey bars indicate the total number of pathways in FEA results. B and C: Top 50 PPs from vote strategy on query LADs of sirolimus and acarbose, respectively. The columns are query samples that are clustered by Euclidian distance of color key (rankings transformed from adjusted p-values). Known LAPs are annotated in green in the binary color bar to the right of the heatmap. The complete tables containing all Reactome pathway rankings from voting strategy with scores and LAP annotation corresponding to the heatmaps are stored in Synapse (syn27074585). D: DT network of the signaling by VEGF (R-HSA-194138) reactome pathway. Symbols of drugs, targets and their relationship are available at Table 3.7.	87
3.7	Combined PPs from LAD11 queries with at least 3 LADs support. The top 50 PPs from each LAD of LAD11 are combined into one final list ranked by the number of query LADs that have the pathway in its top 50 PPs (Nsupport) and filtered by selecting those that have at least 3 Nsupport. It results in 74 Reactome pathways in this list. The corresponding table with no Nsupport cutoff is at Table S11.	88
3.8	Peters <i>et al.</i> (2015) query results. A: NCS score distributions in the GESS results from Peters <i>et al.</i> (2015) query and one random GES query as negative control after setting count cutoff as 100, the left and right panels show the negative part and positive part of NCS scores, respectively. The color key shows the P-values of the WTCS score for the entries in the GESS results. B: Position of known LADs from DrugAge database in Peters GESS result. Due to space limitation, it only shows the position of LADs in the top 500 drugs whose GESs are positively connected with the query GES out of 8140. C: DT network for the EGFR tyrosine kinase inhibitor resistance (EGFR-TKIR) pathway. Symbols of drugs, targets and their relationship are available at Table 3.12.	94
3.9	Distribution of ranking moving average on MOA sizes using moving window analysis. Ndrugs: number of drugs in MOAs, i.e. MOA size. RankMA: moving average of MOA rankings where MOAs ranked from lowest to largest by their sizes. The moving average of their rankings were calculate with a window size of 50. Since a lot of MOAs have size less than 4 and the RankMA was calculated at each MOA, the dots at small Ndrugs are vertical aligned.	109

3.10	Illustration of MOA recall method. Drug GESs of each MOA category were iteratively queried against the LINCS expression database with the <i>SPsub</i> method. All MOAs were ranked from largest to lowest by median absolute CORct of of their drugs across drug GES queries of the query MOA. The rank percentile of the query MOA was set as recall rate.	111
3.11	Permutation testing method. Query GESs were randomized at 5% , 10% or 15% of the GES query genes by iteratively replaced with randomly selected genes from the reference database 1000 times prior to performing the GESS/FEA analysis. The permutation results were compared to the original result rankings and the permutation/robustness scores were calculated as the percent of permutation times that the drug/pathway rankings of the permutation results within the +2 and -2 ranking windows of the non randomized query GESS results.	114
3.12	Illustration of the NCStissue score (A), voting strategy (B) and classification (C). The compounds' NCS scores in cell lines that matches the query tissue (primary cell has first priority and then immortal cell line) are selected. The NCS scores in other cell lines are ignored. The maximum NCS scores are selected for compounds that do not have treatments in the matched cell lines. The cell type summarized GESS results from different GESs of one query LAD are combined into one table. The rank-transformed table is turned into TRUE or FALSE values by setting the rank cutoff (e.g. 400). The compounds in GESS results are then summarized across queries by ranking from largest to lowest by row sums of cell type summarized NCS scores after cutoff. The drug rankings from voting strategy are then classified into three layers (layer 1: drugs matching MOAs with query LADs; layer 2: drugs matching target pathways with query LADs; layer 3: others). Drugs in each layer remain the original voting order. The top ranking drugs in each layer are selected as the final prioritized drugs.	117
3.13	Voting strategy on FEA results. Each of the FEA results from the GESs of one query LAD was ranked from lowest to largest by the adjusted p-values and combined into one table. The rank cutoff of 100 was applied. The pathways in FEA results were summarized across queries by ranking from lowest to largest by row means of rankings after cutoff multiplied with a size correction factor. The latter was calculated by multiplying the total number of GES queries divided by the number of supported queries after rank cutoff to give it a larger value to pathways that are supported by fewer queries to make it rank lower, thus favoring the pathways that are supported by more queries.	119

4.1	Screenshot of the <i>geneTargetAnno</i> web interface that shows the gene target annotation results in DrugBank database for the query Ensembl gene ID. The main utilities were marked in red symbols. The check marks (number 4) show the columns in the DrugBank annotation table including UniProt ID of the query gene (Uniprot_id), DrugBank IDs of the target drugs (target_drugs), drug structures (structure) and UniProt IDs of the drug targets (drug_targets). The brackets in the drug_targets column also include the supported species of the drug-target interaction.	146
4.2	Screenshot of the <i>geneTargetAnno</i> web interface that shows the gene target annotation results in the STITCH database. The columns in the STITCH annotation table include the Ensembl protein IDs of the input Ensembl gene ID (ensembl_protein_id), PubChem CIDs of the target drugs (target_drugs), drug structures (structure) and Ensembl protein IDs of drug targets (drug_targets)	147
4.3	Full gene-target annotation result table for the DrugBank database. It can be searched at the search box (number 1). Each column can be ranked alphabetically (number 2) or searched (number 3). The full table can be downloaded as csv, Excel or pdf file via the buttons at number 4.	148
4.4	Structures of the three out of six compounds in Table 4.4 that have the corresponding SDF instances.	152

List of Tables

2.1	Categories of GESS algorithms by data types. The table compares the different data types used as queries and databases by the GESS methods implemented in <i>signatureSearch</i> . The specific GEP types used by the methods are: ^a rank transformed profiles, ^b Z-scores, ^c normalized intensities or read counts. ^d Pearson or Spearman correlation coefficient.	20
2.2	List of important functionalities provided by <i>signatureSearch</i> and <i>signatureSearchData</i> . The names of functions and libraries are italicized. ^a Only the most common input types are listed. Acronyms are defined in the text.	37
2.3	GESS methods applied to MOA categories. To assess whether the observed performance differences are statistically significant for all pair-wise comparisons, the bootstrap method was used for both the global AUC and pAUC metrics. The BH method was used for multiple testing correction [12]. The columns contain: GESS method 1 ^a ; GESS method 2 ^b ; P-value ^c and adjusted P-value ^d	41
2.4	GESS methods applied to SSC categories. The column titles and content of this table are organized the same way as in Table 2.3.	42
2.5	Time and memory performance	47
2.6	Top ranking drugs of vorinostat query. The GES of SKB cells treated with vorinostat was used as query to search the LINCS database with the <i>SPsub</i> method. The rows are sorted decreasingly by absolute Spearman Correlation Coefficients ^d . The other columns include ranks ^a , drug names ^b , cell types ^c , and the gene symbols of the corresponding target sites ^e	50
2.7	Top ranking MF and BP terms obtained with <i>dup_hyperG</i> . The columns contain: GO Ontology ^a ; GO Term description/ID ^b ; number of proteins in GO term ^c , test set ^d and intersect ^e , raw p-value ^f , and adjusted p-value ^g using the BH method for multiple testing correction. To save space, longer GO term descriptions have been shortened.	52

2.8	Top ranking GO MF and BP terms obtained from direct enrichment of the vorinostat GS-Q with hypergeometric test. The columns contain: GO Ontology ^a ; GO Term description/ID ^b ; number of genes in GO term ^c , test set ^d and intersect ^e , respectively; as well as enrichment p-value ^f and adjusted p-value ^g using the Benjamini-Hochberg (BH) method. To save space, longer GO term descriptions have been shortened.	52
3.1	Overview of 11 high-confidence LADs. LAD data sources are: ITP ¹ : Interventions Testing Program from National Institute on Aging (NIA); Fuentealba <i>et al.</i> (2019) ² ; Strong <i>et al.</i> (2016) ³ . Additional annotation information is available in Table S1.	62
3.2	Top 10 ranking drug cell terms in the GESS result from one sirolimus GES query in SKB cell with LINCS GESS method. RRS5: rank robustness score at 5 percent randomization.	78
3.3	Top 10 ranking Reactome pathways in the FEA result from one sirolimus GES query in SKB cell with <i>dup_hyperG</i> method. N term/t/m: number of genes in the pathway, test set and intersect. P-adjust: P-value using the Benjamini-Hochberg (BH) method for multiple testing correction.	81
3.4	Summary table of LAD11 query GESs from in vitro and in vivo across different cell types or tissues. The starred cell types are used when the LAD names are used as sample names in the results. Condition: mouse condition when sacrificed, the numbers are age in month, M/F represent female/male, chronic means longer term treatment on old mouse.	82
3.5	The manually curated annotations for all of the prioritized DrugAge drugs (PriODA) in this project including drug description, tested assays, publications, <i>etc.</i> PCID: PubChem CID. The elaborate information is at Table S16.	83
3.6	Overlapped Reactome pathways between top 50 enriched terms from Sirolimus top 50 prioritized drug list and top 50 prioritized list in Sirolimus FEA results.	89
3.7	Drugs and targets involved in the DT network of signaling by VEGF (R-HSA-194138) in Figure 3.6D.	90
3.8	Top 10 ranking positively connected drugs of Peters GESS result from <i>LINCS</i> method. Cell: cell type, NEU: normal cell, PC3: prostate adenocarcinoma, VCAP: prostate carcinoma, ASC: normal primary adipocyte stem cells; Targets: gene symbol of protein targets; RRS5,10,15: rank robustness scores at 5, 10, 15 percent randomization. For detailed description of the score columns for LINCS method, please consult to the vignette of the <i>signatureSearch</i> package. The complete table is available at Table S13.	94
3.9	Overlapped drugs between top 500 positively connected drugs in Peters GESS results and combined PDs summarized from LAD11. Therapeutic: whether the drugs have therapeutic effect. Max Phase: FDA max phase study.	95
3.10	Top 10 ranking KEGG pathways from Peters <i>et al.</i> (2015) longevity-based query GES. RRS5, 10, 15: rank robustness scores at 5, 10 and 15 percent randomization. The complete table is available at Table S14.	96
3.11	Overlapped Reactome pathways between top 100 terms in Peters FEA results and combined PPs summarized from LAD11.	97

3.12	Summary of the NCS scores, MOA, target gene symbols and targets in network of the filtered 11 drugs in the EGFR-TKIR drug-target network.	98
4.1	Six drug entries in DrugBank database with several selected column annotations as an example.	144
4.2	Annotation table of the name of the query DrugBank IDs and whether they are FDA approved with logic values.	144
4.3	Target annotation table of the query drugs. The queries are DrugBank IDs (Q-DBID). Their gene/protein targets have four ID systems including DrugBank target ID (T-DBID), UniProt ID (T-UnipID), UniProt name (T-UnipName) and gene symbol (T-Gene). For queries that have many targets, only three out of them are shown. The others are deleted and replaced with three dots for better display.	145
4.4	Top six rows of the DrugAge annotation table with several selected columns. Avg: average lifespan change, PCID: PubChem CID, DBID: DrugBank ID.	152
4.5	Annotation table of the five query compounds with ChEMBL IDs in LINCS and custom databases. isTS: whether the compounds are in Touchstone database, PCID: PubChem CID, feature1, 2: two columns in the custom annotation table added by user.	154

Chapter 1

Introduction

1.1 Overview

Genome-wide profiling technologies for mRNAs and proteins provide comprehensive recordings of biological processes. Their high-resolution can be used to distinguish cell and tissue types, and to classify dynamic cellular processes into distinct biological states such as developmental stages, defense responses to perturbagens, as well as to separate healthy from diseased phenotypes [109, 173]. To take full advantage of the fingerprint-level selectivity of the technology, so called Gene Expression Signature (GES) Search (GESS) algorithms are essential to accurately quantify the similarities among mRNA or protein profiles available in reference databases. With these methods one can identify similar GESs that are likely to be induced by the same or related biological mechanisms [162]. This approach is analogous to the *similarity-function principle* used in many areas of biology, such as in genomics where genes with a high degree of sequence similarity are likely to share similar molecular functions.

With the availability of databases containing GESs of thousands of treatments tested on many cell types, it is now possible to systematically search for genetic backgrounds, diseases, physiological conditions or small molecules inducing gene expression responses that are similar to a query GES. Both positively or negatively correlated search hits can provide insights into previously unknown connections among biological networks. For example, distinct diseases may lead to overlapping mRNA expression patterns resulting from the same or related immune response processes. Mutations inducing similar GESs may allow to functionally associate them with biological processes even if the affected genes do not share detectable sequence similarities. Similarly, drugs used for different therapeutic applications may have similar GESs due to related mode of actions (MOAs). Among other leads, this information can be used for identifying novel drug targets or for developing drug repurposing approaches [34]. Ultimately, the technology has the potential to lead to the discovery of novel pharmaceutical treatments for diseases, such as for health conditions characterized by specific GESs that are anti-correlated with those of candidate drugs. Beyond these utilities, the GESS technology has a wide application spectrum for addressing fundamental research problems in biology and human health.

An important requirement for the GESS technology is the availability of reference databases containing GESs suitable for addressing specific research questions. GESs can be composed of gene sets (GSs), such as the identifier sets of differentially expressed genes (DEGs), or various types of quantitative gene expression profiles (GEPs) for a subset or all genes measured by a gene expression profiling technology. Some publications refer with the term GES mainly to GSs, or use as extended terminology ‘qualitative and quantitative

GESs' [28]. For clarity and consistency, I defines GES as a generic term that comprises both GSs and GEPs [109]. This generalization is important, because several GESS algorithms are introduced here that depend on reference databases containing GSs in some and GEPs in the majority of cases generated with various statistical methods. To also distinguish the queries (Q) from the entries in the reference databases (DB), they will be referred to as GES-Q and GES-DB entries in general descriptions, and as GS-Q or GEP-Q, and as GS-DB or GEP-DB in specific cases, respectively.

Three major approaches are commonly used to assemble community GES-DBs. First, they can be assembled from the results of published genome-wide expression experiments. Due to the heterogeneous nature of how result tables are organized in publications, the corresponding publication-based collections are often composed of GSs (*e.g.* DEGs in GS-DBs). Examples in this category include GeneSigDB, MSigDB, DSigDB and GSKB [22, 117, 36, 107, 228]. Second, both GS-DBs and GEP-DBs have been assembled by systematically re-analyzing genome-wide expression data from public repositories such as GEO [201, 136]. This reanalysis approach allows to include the corresponding numeric expression data, while also using consistent statistical methods for normalization, DEG detection and other analysis routines across studies. Third, large-scale experimental screening efforts have been used to assemble GEP-DBs, such as for a wide range of genetic and drug perturbation measurements across many cell types. These *de novo* screening efforts allow a high level of control over both experimental conditions as well as statistical analysis methods. Specific examples of GEP-DBs are described in the next paragraph. Importantly, all three categories of GES-DBs are supported by the GESS methods.

One of the first screening-based GEP-DBs [86] was developed by Hughes et al. (2000). It contained GEPs of 300 diverse mutations and chemical treatments to functionally annotate both small molecules and genes in yeast. The study demonstrated that the cellular pathways perturbed by genetic modifications or small molecules can be determined by pattern matching. In mammalian biology, Ganter *et al.* (2005) generated a large-scale GEP-DB containing perturbations of several rat tissues with 600 drugs [59]. They also demonstrated the utility of GEP-DBs for predicting pathological events in rats. However, these *in vivo* studies did not easily scale to larger quantities of small molecule assays mainly due to the high cost and time of performing compound screens on living animals. Lamb *et al.* (2006) generated the first large-scale mammalian cell line-based GEP-DB, called ‘Connectivity Map’ or CMAP [109]. Initially, it included GEPs for 164 drugs screened against four mammalian cell lines [140]. A few years later CMAP was extended to CMAP2, which contains GEPs for 1,309 drugs and eight cell lines. More recently, a much larger GEP-DB was released by the Library of Network-Based Cellular Signatures (LINCS) Consortium [186]. In its initial release, the LINCS database contained perturbation-based GEPs for 19,811 drugs tested on up to 70 cancer and non-cancer cell lines along with genetic perturbation experiments for several thousand genes. The number of compound dosages and time points considered in the assays has also been increased by 10-20 fold. The CMAP/CMAP2 databases use Affymetrix Gene Chips as the platform for expression analysis. To scale from a few thousand to many hundred thousand GEPs, the LINCS Consortium uses the more economic L1000 assay. This bead-based technology is a low cost, high-throughput reduced representation expression profiling assay. It measures the expression of 978 landmark genes and 80 control genes by

detecting fluorescent intensity of beads after capturing the ligation-mediated amplification products of mRNAs [140]. The expression of 11,350 additional genes is imputed from the landmark genes by using as training data a collection of 12,063 Affymetrix gene chips [37]. The substantial scale-up of the LINCS project provides many new opportunities to explore MOAs for a large number of known drugs and experimental drug-like small molecules. Complementary proteomics GES-DBs are also being developed by several community projects [50]. Additional large-scale expression data and databases, where GESS applications can lead to interesting findings, consider cancer, tissue-specific, and single cell assays, such as TCGA, GTex and Single Cell Portal, respectively [21, 69, 1].

Because GESS results are usually composed of complex lists of perturbagens (*e.g.* drugs) ranked by their GES similarity to a GES-Q of interest, their functional interpretation is difficult with respect to the cellular networks and pathways affected by the top ranking results. In the case of drug-based GES-DBs, one can overcome this challenge by utilizing the knowledge of the target proteins of the top ranking drugs to perform functional enrichment analysis (FEA) based on community annotation systems, such as Gene Ontology (GO), pathways (*e.g.* KEGG, Reactome), drug MOAs, or Pfam domains. To perform this analysis, the ranked drug sets are converted into the corresponding target gene/protein sets they modulate, and then Target Set Enrichment Analysis (TSEA) based on a chosen functional annotation system is pursued. Alternatively, the functional annotation categories of the targets can be assigned to the drugs directly to perform Drug Set Enrichment Analysis (DSEA).

1.2 Need to Develop GESS Software Environment

Currently, no one-stop software solution is available to perform the analyses outlined above in an integrated manner using a variety of GESS/FEA algorithms across several pre-built or custom GES-DBs. Previous work in this field includes web-based tools [43, 186, 109, 181, 147] and standalone software [111, 162]. Both types are usually restricted to the usage of specific pre-configured GES-DBs of limited size with insufficient options to choose among GESS methods. To address these limitations, I have developed the *signatureSearch* software. This R/Bioconductor package provides several important enhancements to the field including access to: (a) an integrated and flexible analysis environment for GESS applications; (b) a wide range of GESS methods; (c) novel enrichment algorithms for interpreting GESS results; (d) data containers, classes and accessor methods designed to scale to very large GES data sets; (e) batch query support for large-scale applications; (f) access to several large pre-built GES-DBs; as well as (g) support for searching custom GES-DBs.

A substantial amount of development effort has been invested by this project to provide efficient access to some of the largest GES-DBs that are currently available in the public domain (*e.g.* CMAP2 and LINCS). Since those databases are designed around chemical perturbation experiments, this project will mainly apply the *signatureSearch* in the drug discovery field. In this context it is important to emphasize that the design of *signatureSearch* is highly generic, meaning it can be used for GES-Qs and GES-DBs from many other research areas in biology or human health.

1.3 Discovery of Healthy Aging Drugs

The GESS/FEA environment implemented in the *signatureSearch* software can be applied to the human longevity field to discover novel longevity associated drugs. Human longevity and healthy aging is a major worldwide medical challenge. It is difficult to identify drugs and compounds that extend human lifespan since human longevity and healthy aging are interconnected and multifaceted phenotypes involving sex differences, genetic, epigenetic, environmental, and lifestyle components. Large epidemiological and genetic studies at the national and international levels have identified genetic markers that are both correlated, and in more isolated cases, causally linked to molecular mechanisms promoting longevity. Genetic risk factors for developing aging-related diseases include mutations in BRCA1, TP53, KRAS and EGFR, whereas lifespan promoting genetic changes comprise ablation of IGF-1, GHR, PAPP-A, IRS2, apolipoprotein E (APOE) and Forkhead Box O 3A (FOXO3A) [125, 183, 7, 163, 32, 213]. Certain epigenetic modifications, such as DNA methylation, are also linked to lifespan increases and aging related diseases [75, 94, 108, 189]. Studies have shown that siblings of nonagenarians [209] and of centenarians [143, 214, 164] have a high probability of living nearly 100 years, lower mortality rate and delayed onset of age-related diseases [191, 144]. Apart from genetic factors, several studies have demonstrated that longevity is amendable to diet and exercise including calorie restriction in the absence of nutrient deprivation, intermittent fasting, as well as restriction of protein, methionine and tryptophan [53]. Further, a higher than normal concentration of centenarians live in so called blue zones of the world. Most likely, the average lifespan within these blue zones has increased over the past decades mainly because of specific lifestyle choices and environmen-

tal conditions, rather than genetic factors [127, 19, 35]. The important role genes play on human healthy aging and the discovery that lifespan can be extended via both behavioral and dietary modifications demonstrate the possibility of identifying functional pathways related to healthy aging and developing broadly applicable therapeutic interventions for the prevention of aging-related disorders [53, 87, 165].

Additional analyses have delineated a number of longevity associated pathways (LAPs) including energy metabolism, cholesterol and lipid metabolism, inflammation, DNA damage response and repair [39], telomere length maintenance [3], and heat shock response [160]. In addition, researchers have found that signaling pathways involved in detecting and interpreting nutrient or energy levels, such as the insulin/insulin-like growth factor 1 (IGF-1) signaling pathway [213, 95], mechanistic target of rapamycin (mTOR) [93], and adenosine monophosphate-activated protein kinase (AMPK) [53, 190] play important roles in regulating transcriptional responses associated with extended reproduction, growth, and lifespan. Clinical studies also demonstrated that genes implicated in lipoprotein metabolism, cardiovascular homeostasis, immunity, and inflammation may have a strong impact on aging, age-related diseases, and organism longevity [163, 24, 40]. Thus far, there has been only limited success translating these findings into life span enhancing pharmacological interventions. In part this is due to a limited ability of identifying compounds that elicit changes in gene expression similar to those involved in longevity.

To date, many pharmaceutical agents have been identified that target LAPs and may have positive effects on lifespan. Most of the well-studied drugs or small molecules that can be used to extend life span in model organisms are populated and curated in

the DrugAge [6] database, which is manually curated by experts and contains an extensive compilation of drugs, compounds and supplements (including natural products and nutraceuticals) with anti-aging properties. At the time of writing, DrugAge contains 1,832 assays manually curated from 469 publications featuring 567 different life-span extending compounds from studies across 30 model organisms, including worms, flies, yeast and mice (<http://genomics.senescence.info/drugs/>). It can be used as a reference database to further identify and prioritize candidate longevity associated drugs (LADs). The drugs that are listed as compounds in testing (CIT) in longevity assays from the Interventions Testing Program (ITP) at the National Institute on Aging [206] can also be a LAD resource. Among these agents are tanespimycin, minocycline, acarbose, resveratrol, aspirin, curcumin, estradiol, simvastatin, rapamycin, and metformin. They are currently under consideration to be used as pharmacological interventions to prevent and/or reverse aging related diseases [15, 185, 171, 78, 8, 80, 76, 178, 112, 57]. Rapamycin (also known as Sirolimus) is an FDA approved drug for preventing organ transplant rejection and treating a rare lung disease with its immunosuppressant activity in humans. It inhibits the mTOR pathway, slows the aging process and extends lifespan in mice, yeast, worms, and flies [92]. However, systematic studies are missing to determine which FDA approved and novel medications are promising candidates for translational lifespan promoting interventions in humans.

The GESS technology can be used to various applications for designing novel health- and lifespan- promoting interventions. The effectiveness of the approach has been demonstrated in a wide range of biological areas [109, 227, 198, 205, 224, 102, 174]. Additionally, the technology has been applied to questions related to aging and longevity in mice

and humans [158, 195, 89]. However, the application of GESS for systematic identification of compounds with potential lifespan extending properties has not been performed yet. To address the knowledge gap, this study used GESS analyses to characterize drugs based on GES similarities followed by FEA using known molecular targets of the top ranking compounds to identify and functionally characterize new and existing LADs and LAPs.

1.4 Integration of Annotations from Other Sources

As extensions of the *signatureSearch* package, I developed several data packages that incorporate detailed annotations and structures of drugs to facilitate the interpretation of GESS results. The chosen annotations focus on drug-target information from different community databases, including DrugBank, DrugAge, CMAP2 and LINCS. Databases such as PubChem [101], ChEBI [79], ChEMBL [61], and DrugBank [215] provide nomenclature, structure and/or physical properties of large numbers of compounds and their drug targets. One of the databases used for extending *signatureSearch* is DrugBank. It covers sequence, structure and mechanistic data about drug molecules with their gene/protein targets [216]. It also contains information about clinical and drug repurposing trials [215]. At the time of writing this thesis, the DrugBank database contained a total number of 14,549 drugs, including 11,898 small molecules, 2,651 biotech drugs, and 4,167 approved drugs (<https://go.drugbank.com/stats>). Another database included here is CMAP2. It provides information about the experimental design of bioassay results of drugs, their names, identifiers and SMILES strings. Additional annotations such as Mechanism of Actions (MOAs) were obtained from PubChem and DrugBank. Drug annotation from LINCS

were also integrated. This provides for over 20,000 small molecules screened with GES data from LINCS their names, SMILES strings, targets, MOAs, *etc.* The annotation and structure data from these sources were organized in an SQLite database and combined with efficient accessor functions to access the data from R. The resulting R/Bioconductor data package was named *customCMPdb*.

To obtain via a web interface drug-target annotations for any gene or protein list provided by users, I developed the Shiny web application *geneTargetAnno*. If the gene ids in the users' provided gene table are included in the application's url, the gene ids can be easily linked to application's drug-target annotation page to retrieve the targeted drugs and their structures for the query genes. The provided gene table can be from any type of biological studies, such as the differential expressed (DE) genes that are from DE analysis on the biological state of interest, or genes affected by single nucleotide polymorphisms (SNPs) prioritized by a genome-wide association study (GWAS) project. It can be used to identify the directly targeted drugs for genes of interest, thus to further identify candidate drugs and potential treatments of diseases if dysregulation of the input genes are related to the disease status. Shiny is a web application framework for R. It can be used to easily build interactive web apps directly from R (<https://shiny.rstudio.com/>). The developed Shiny apps can be run locally or deployed on web services including AWS and shinyapps.io. The *geneTargetAnno* service integrates annotations about drug-target interaction, tested organisms and structure information from DrugBank and STITCH (Search Tool for Interacting Chemicals). The latter contains protein-chemical interaction networks obtained from databases and prediction methods [188]. The most recent release of the STITCH database contained

data from more than 9,600,000 proteins of 2,031 eukaryotic and prokaryotic genomes, and 430,000 compounds (<http://stitch.embl.de>).

The R/Bioconductor packages and Shiny web service developed by this project provide several important enhancements to the cheminformatic field including (a) access to a pre-configured SQLite database containing is a comprehensive collection of compound annotations from DrugBank, DrugAge, CMAP2 and LINCS; (b) query functions for drug-target analysis as well as obtaining detailed compound annotations; (c) compatibility with other cheminformatics tools such as *ChemmineR* for a variety of compound rendering, similarity search and clustering methods, as well as (d) data containers, classes and accessor methods designed to add and query custom compound annotations.

Chapter 2

signatureSearch Tool

2.1 Abstract

signatureSearch is an R/Bioconductor package that integrates a suite of existing and novel algorithms into an analysis environment for gene expression signature (GES) searching combined with functional enrichment analysis (FEA) and visualization methods to facilitate the interpretation of the search results. In a typical GES search (GESS), a query GES is searched against a database of GESs obtained from large numbers of measurements, such as different genetic backgrounds, disease states and drug perturbations. Database matches sharing correlated signatures with the query indicate related cellular responses frequently governed by connected mechanisms, such as drugs mimicking the expression responses of a disease. To identify which processes are predominantly modulated in the GESS results, I developed specialized FEA methods combined with drug-target network visualization tools. The provided analysis tools are useful for studying the effects of genetic, chemical and environmental perturbations on biological systems, as well as searching single cell GES

databases to identify novel network connections or cell types. The *signatureSearch* software is unique in that it provides access to an integrated environment for GESS/FEA routines that includes several novel search and enrichment methods, efficient data structures, and access to pre-built GES databases, and allowing users to work with custom databases.

2.2 Materials and Methods

2.2.1 Implementation

signatureSearch has been implemented as an open-source Bioconductor package using the R programming language for statistical computing and graphics. The affiliated data package *signatureSearchData*, provides direct access to large data sets, such as pre-built GES-DBs and annotation databases that are hosted on Bioconductor's ExperimentHub. Both packages are freely available for all common operating systems. To optimize reusability and performance, their functions and data containers are designed based on existing Bioconductor S4 core classes. Some of the time consuming computations have been implemented in C++ using R's C++ interface. Additional implementation details are provided in the *Software Design* section below. Up-to-date source locations and versions of data sets are provided in the vignettes and help files of the two packages.

2.2.2 Data Types of Queries and Databases

As outlined in the Introduction section, GESs of both queries and those stored in databases can be composed of GSs, or various types of quantitative GEPs for all genes measured by a gene expression technology or only a subset of them. Depending on the extent

the expression data have been pre-processed, the following distinguishes four major levels, where the first three and fourth belong into the GEP and GS categories, respectively. These four levels are: (1) normalized intensity or count values from hybridization- and sequencing-based technologies, respectively; (2) log fold changes (LFC) usually with base 2, Z-scores or p-values obtained from analysis routines of DEGs; (3) rank transformed versions of the GEPs obtained from the results of level 1 or 2; and (4) GSs extracted from the highest and lowest ranks under level 3. Typically, the corresponding GSs are the most up- or down-regulated DEGs observed among two biological states, such as comparisons among untreated *vs.* drug treatment or disease state. The order the DEG identifier labels are stored may reflect their ranks or have no meaning. When unclear, the text specifies which of the four pre-processing levels were used along with additional relevant details.

2.2.3 Reference Databases

The GESS algorithms and data structures provided by *signatureSearch* and *signatureSearchData*, respectively, are designed to work with most genome-wide expression data including hybridization- and sequencing-based methods, such as Affymetrix or L1000, and RNA-Seq. Currently, the pre-built GES-DBs in *signatureSearchData* include GEP data from the CMap and LINCS projects that are largely based on drug and genetic perturbation experiments performed on variable numbers of human cell lines [109, 186, 47]. The CMap data were downloaded from the CMap project site (Version build02), and the LINCS data have been downloaded from GEO. Additional details on the content and design of these databases are provided in the Introduction of this article. In *signatureSearchData* these data sets have been pre-processed to be compatible with the different GESS algorithms implemented in

signatureSearch (Table 2.1). Additional details about pre-processing routines are available in the following section as well as the package documentation. In addition, the package provides functions along with user instructions for generating custom databases that are compatible with the corresponding GESS methods. Moreover, instructions are provided how to work with other public domain GES-DBs including GS-DBs, such as MSigDB and GSKB [22, 117, 36, 107, 228].

2.2.4 Pre-processing and cutoffs for queries and databases

The quantitative gene expression profiling (GEP) data used by this study were downloaded from Bioconductor’s ExperimentHub with utilities provided by the *signatureSearchData* package. The latter provides pre-configured data sets for this project. At the time of writing, the GEP databases (GEP-DBs) included in *signatureSearchData* are LINCS and CMAP2 [109, 186]. Since the experiment section of this article uses LINCS data, the following focuses on the pre-processing and filtering routines of this dataset. The non-quantitative gene sets (GSs) used as GS queries (GS-Qs) and GS databases (GS-DBs) in the article were also extracted from LINCS. The corresponding filtering parameters for obtaining these GSs are given in the next paragraph. Similar information, with additional details for both LINCS and CMAP2, is available in the vignette of the *signatureSearchData* package. Although CMAP2 was not used in the experiment section of this article, the following does include an overview of the corresponding pre-processing routines of this data set mainly to illustrate how to use CMAP2 instead of LINCS for similar analyses.

LINCS GEP data

The Broad Institute has generated the LINCS GEP data with the bead-based L1000 assay for gene expression profiling. Since this technology is not widely used yet and pre-processing methodologies for its data are limited in the public domain, I have chosen to use the pre-generated data instances from the LINCS project directly rather than attempting to regenerate them from raw data. The GEP data from LINCS data can be downloaded from GEO in 5 different pre-processing levels [186]. Level 1 data are the raw mean fluorescent intensity values that come directly from the Luminex scanner. Level 2 data are the expression intensities of the 978 landmark genes. They have been normalized and used to impute the expression of an additional set of 11,350 genes, forming Level 3 data. A robust Z-scoring procedure was used to generate differential expression values from the normalized profiles (Level 4). Finally, a moderated Z-scoring procedure was applied to the replicated samples of each experiment (mostly 3 replicates) to compute a weighted average signature (Level 5). For a more detailed description of LINCS' pre-processing methods, readers want to refer to the methods section in the corresponding publication by Subramanian *et al.*, 2017 [186].

The differential expression data from LINCS used in this article are level 5 Z-scores. Since some GESS methods such as *gCMAP* and *Fisher* require gene sets in the reference database, Z-score cutoffs can be used to filter for sets of up- and down-regulated differentially expressed genes (DEGs). In this article, the corresponding up or down DEG sets were obtained with Z-score cutoffs of ≥ 1 or ≤ -1 , respectively. In *signatureSearch*, these Z-score cutoffs can be assigned to filtering arguments to generate either query or database instances meeting the corresponding Z-score constraints. Examples of GS-DBs

where this is relevant are those used by the *gCMAP* and *Fisher* GESS methods. In addition to using Z-score cutoffs, GS-Qs can also be extracted by specifying a fixed number of the most extremely up- and down-regulated genes, such as the top 150 up- and down-regulated DEGs, respectively. Whether GS-Qs or GS-DBs instances were obtained by Z-score or number cutoffs is specified in the corresponding sections of the article. If the cutoff parameters deviate from the above default values then they are given as well. Examples of GESS function calls related to these routines are provided in the vignettes of the software and data packages of the *signatureSearch* environment. For instance, the subsection ‘*DEG and Cutoff Definitions*’ in the *signatureSearchData* vignette provides details on this topic.

CMAP2 GEP data

This section provides a short overview of the CMAP2 data pre-processing steps to illustrate how this drug-perturbation GEP-DB could be used instead of LINCS for the performance test and proof-of-concept experiments included in this article. Both databases are supported by *signatureSearchData*, but for consistency I only used the LINCS database in the experimental sections. Since the Affymetrix GeneChip[®] technology used by CMAP2 is supported by a rich ecosystem of widely used analysis software, I generated the pre-processed and final data tables for this data set from the corresponding raw files (here CEL files), and deposited the results on Bioconductor’s ExperimentHub for easy access with *signatureSearchData*. To compare the search results generated with the CMAP2 online service and the GESS methods from *signatureSearch*, I also included the CMAP2 rank matrix that is based on rank transformed differential expression values for all assayed genes. The latter can be downloaded from the CMAP2 project site. For the raw data processing

from CEL files, normalized gene expression data were generated with the MAS5 algorithm [142]. Next, the DEG analysis was performed with the *limma* package [154] using the experimental design table included in the CMAP2 data set to define replicates, as well as control and treatment samples. The statistical result tables generated by *limma*, including LFC values, p-values and false discovery rates (FDR), were saved to the HDF5 files I am hosting on Bioconductor’s ExperimentHub. These statistical values can be used by the query retrieval and GESS methods in *signatureSearch* to define DEGs with single or combinatorial cutoff parameters, such as DEGs that have an LFC value of ≥ 1 or ≤ -1 , and an FDR of ≤ 0.01 . Although the LINCS and CMAP2 result tables had to be generated with different statistical methods, one can filter in both cases for DEGs with cutoffs that can be applied to statistical values with comparable meaning (*e.g.* LFCs can be used instead of Z-scores). Detailed instructions along with the corresponding R code for creating the corresponding gene expression and statistical result tables are provided in the CMAP2 pre-processing sections of the *signatureSearchData* vignette. For instance, instructions for defining DEG sets with combinatorial filters of statistical parameters are given in the Supplement section of the vignette under ‘*DEG and Cutoff Definitions*’.

2.2.5 Compatibility Among Data Types

The types of query and database GESs that can be combined in a search usually depends on the chosen GESS algorithm. To avoid incorrect selections for users, the corresponding GESS functions in *signatureSearch* enforce the usage of compatible query and database combinations. Which GES types are compatible with each search method is summarized in Table 2.1.

Table 2.1: Categories of GESS algorithms by data types. The table compares the different data types used as queries and databases by the GESS methods implemented in *signatureSearch*. The specific GEP types used by the methods are: ^arank transformed profiles, ^bZ-scores, ^cnormalized intensities or read counts. ^dPearson or Spearman correlation coefficient.

Category	Method	Query	Database
Set-based	CMAP	GS	Rank ^a
	gCMAP	Rank	GS
	LINCS	GS	Z-scores ^b
	Fisher exact	GS	GS
Correlation	PCC/SCC ^d	LFC or SIG ^c	LFC or SIG

2.2.6 Overview of Analysis Workflow

A typical analysis workflow in *signatureSearch* consists of three major steps (Figure 2.1). First, GESS methods are used to identify biological states or perturbagens such as drugs that induce GESs similar to a query GES of interest. The queries can be GSs or GEPs from genetic, drug or disease perturbations, as well as from many other experiment types. When using a GES-DB based on drug perturbations such as LINCS, then the MOAs of most drugs represented by GESs in the corresponding reference databases are known. With this information one can associate a query GES with the corresponding molecular mechanisms including available drug-target interactions. The obtained connections are useful to gain insights into pharmacological and/or disease mechanisms, and to develop novel drug repurposing approaches. Second, specialized functional enrichment analysis (FEA) methods using annotations systems, such as Gene Ontology (GO), pathways or Disease Ontology (DO), have been developed and implemented in this package to efficiently interpret GESS results. The latter are usually composed of lists of perturbagens (*e.g.* drugs or mutations) ranked by the GES similarity scores returned by a chosen GESS method. Interpreting

these lists of perturbagens without *signatureSearch*'s functional interpretation methods is extremely difficult. Third, network reconstruction functionalities are integrated for visualizing the final results, *e.g.* in form of drug-target networks (DTN). Figure 2.1 illustrates the major steps of a typical workflow in *signatureSearch*. For each GESS and FEA step, several alternative methods have been implemented in *signatureSearch* to allow users to choose the best possible workflow configuration for their research application. Basic guidelines for choosing software tools are provided below as well as in the documentation of the package. The individual search and enrichment methods are introduced in the sections below.

For users working in drug discovery or chemical genomics, a rich suite of cheminformatics functionalities is readily available to enhance the above workflow via the affiliated *ChemmineR* package [23, 203]. This way one can start with structure similarity searches to first identify related drugs represented as perturbagens in a GES database. Subsequently, the corresponding GESs are used as queries in the above GESS/FEA workflow. Moreover, one can cluster GESS results by structural or physicochemical similarities of the corresponding small molecules, *e.g.* to assess the quality of GESS results. The approach is based on the assumption that related compounds are more likely to induce similar GESs resulting in similar GESS rankings.

2.2.7 Analysis Methods

The following describes the methods used within each of the three major steps of a *signatureSearch* analysis workflow.

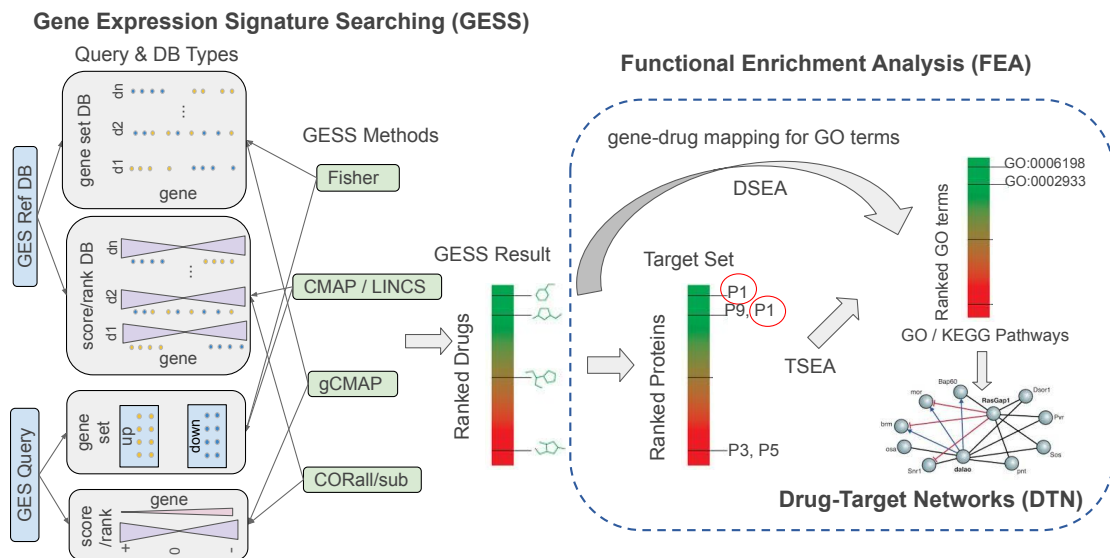


Figure 2.1: Overview of GESS and FEA workflow. GES queries are used to search a drug-based GES reference database for drugs inducing GESs similar to the query. To interpret the results mechanistically, the GESS results are subjected to functional enrichment analysis (FEA) including drug and target set enrichment analyses (DSEA, TSEA). Both identify enriched functional categories (GO terms and/or KEGG pathways) in the GESS results. Subsequently, drug-target networks (DTNs) are reconstructed for visualization and interpretation.

GESS Methods

At the time of writing *signatureSearch* includes five GESS algorithms, with additional algorithms to be added in the future. Alternatively, users can provide their own GESS methods. Based on the data types represented in the query and database, they can be classified into set- and correlation-based methods (see Table 2.1 and Figure 2.1). The first 4 methods described below are set-based, whereas the last one is a correlation-based method. I refer to a search method as set-based if at least one of the two data components (query and/or database) is composed of a GS (*e.g.* gene labels) that may be ranked or unranked. In contrast to this, correlation-based methods require quantitative GEPs, usually of the same type for both the query and the database entries, such as normalized fluorescence intensities, read counts or Z-scores. An advantage of the set-based methods is that their queries can be the highest and lowest ranking gene sets in each direction derived from a genome-wide profiling technology that may differ from the one used to generate the reference database. However, the precision of correlation methods often outperforms set-based methods as will be shown in the Result section. This is most likely a result of the larger information content used by correlation-based methods compared to set-based methods. On the other hand, due to the nature of the expected input, correlation-based methods are usually only an option when both the query and database entries are GEPs generated by the same or at least comparable expression assay technologies. In other words, set-based methods are more technology agnostic than correlation-based methods, but may not provide the best recall performance as shown below. The following describes the most important features of each of the five GESS methods. For clarity and simplicity, the first method will be introduced in

more detail, while the descriptions of the remaining ones will focus mainly on their common and unique features. Since the data types used for the queries and reference databases are different among the GESS methods, the corresponding input requirements will be specified for all of them. This is important both to understand the basic principles of the algorithms, and to choose the appropriate GESS methods for specific data sets available to users.

The Connectivity Map (CMap) GESS method [109], here termed as CMAP, uses as GS-Qs the most strongly up- and down-regulated genes from an experiment, while the reference database is composed of rank transformed GEPs (*e.g.* ranks of LFC or Z-scores) containing all genes or proteins detected by the underlying expression technology. The actual GESS algorithm of the CMAP method is based on identifying a maximum in a vectorized rank difference calculation for each of the up and down GS-Qs separately [109]. After subtracting the down from the up maximum, or assigning zero to certain exceptions, the resulting raw scores are scaled to values from 1 to -1. The final ‘Connectivity Scores’ expresses to what degree the up and down components of the GS-Q are enriched on the top and bottom of each database entry, respectively. The search results are a tabulated representation of the identifiers and descriptions of each GEP entry in the reference database that can be ranked by the connectivity score obtained for the corresponding GS-Q. If the utilized GEP-DB was obtained from drug perturbation experiments then the corresponding GESS scores indicate which drugs induce similar or opposing GESs as the query. Although several variants of the CMAP algorithm are available in other software packages including Bioconductor, the CMAP implementation provided by *signatureSearch* is unique by following the original description of the authors as closely as possible. This allows the reproduction of the

search results obtained from the corresponding CMAP2 web service of the Broad Institute. Determining whether the results generated by both tools will consistently be the same for any GS-Q is not feasible at this point, because CMAP2 is only available as a web service that does not support large-scale queries required for systematic performance testing.

A more complex GESS algorithm was introduced by Subramanian et al. (2017), here referred to as LINCS method. While related to the original CMAP method, there are several important differences among the two approaches. First, LINCS weights the genes in the GS-Q based on the corresponding differential expression values of the GESs in the reference database (e.g. LFC or Z-scores). Thus, the reference database used by LINCS needs to store the actual differential expression values rather than their ranks. Another relevant difference is that the LINCS algorithm uses a bi-directional weighted Kolmogorov-Smirnov enrichment statistic to compute a ‘Weighted Connectivity Score’ (WTCS) as similarity metric. If experimental design groups for the GEP entries in the database are available, such as shared cell types and treatment types, then the WTCS can also be normalized and standardized to obtain the ‘Normalized Connectivity Scores’ (NCS) and ‘Standardized Enrichment Scores’ (τ), respectively. To the best of our knowledge, the LINCS search and scoring functionalities in *signatureSearch* provides the first downloadable standalone software implementation of this algorithm.

The Bioconductor gCMAP [162] package provides access to a related but not identical implementation of the original CMAP algorithm described above. While the computation of the connectivity score is similar, the main difference is that gCMAP uses as a query a rank transformed GEP and each entry in the reference database is a GS composed of the

labels of up- and down-regulated DEG sets. This is the opposite situation of the CMAP method, where the query is composed of the labels of up- and down-regulated DEGs and the database contains rank transformed GEPs.

Fisher’s exact test [68] can also be used as a GESS method by iteratively running the test to assess the degree of similarity shared among a GS-Q with each entry in a reference GS-DB. This method performs an over-representation analysis based on a two-by-two incidence matrix. The latter comprises set comparison counts for each GS comparison pair, including the number of genes in each GS, the numbers of their common and unique genes, the total number of genes in the reference database (universe), as well as certain derivatives of these numbers. The resulting enrichment probabilities are based on the hypergeometric distribution. To account for the multiple hypothesis testing situation of a search result, the obtained p-values are adjusted with the Benjamini & Hochberg method [12]. In this case the search method is entirely set-based, because both the query and the database entries are composed of GSs, such as DEG sets. When the reference database is a quantitative GEP-DB then it can be converted to a GS-DB in *signatureSearch* on the fly using a user-definable cutoff (*e.g.* score or p-value).

If both the query and the database entries are available as numeric GEPs then correlation-based similarity metrics [56], such as Spearman or Pearson correlation coefficients, can be used as GESS methods. In short, correlation methods express the strength and direction of a linear relationship between two sets of paired numeric values (*e.g.* two GEP vectors) with a correlation coefficient. The latter is defined as the covariance of the numeric values divided by the product of their standard deviations. As non-set-based meth-

ods, they require the same type of quantitative gene expression values for both the query and the database entries, such as normalized intensities or read counts from microarrays or RNA-Seq experiments, respectively. The correlation-based searches can either be performed with the full set of genes represented in the database or a subset of them. The latter can be useful to focus the computation for the correlation values on certain genes of interest such as a DEG set or the genes in a pathway of interest. In this regard the correlation-based GESSs, performed on subsets of genes, are unique in one important aspect. That is, they allow generating meaningful GESS results for GEP-Qs, where the corresponding query genes can be derived from a variety of sources or custom collections. This means they are not necessarily expected to be the highest ranking gene or protein candidates, such as DEGs, discovered in a genome-wide profiling experiment as it is often expected for most set-based methods. The following refers to a correlation-based GESS as *SPall* or *SPsub* when considering in a search with the Spearman method the data of all assayed genes or only a subset of them (*e.g.* DEG set), respectively.

FEA Methods

GESS results are lists of GEP-DB or GS-DB entries ranked by the similarity metric of a chosen GESS method. When searching drug-based GES-DBs, then the corresponding drugs are ranked accordingly. Interpreting these search results with respect to the cellular networks and pathways affected by the top ranking drugs is difficult. To overcome this challenge, the knowledge of the target proteins of the top ranking drugs can be used to perform functional enrichment analysis (FEA) based on community annotation systems, such as Gene Ontology (GO), pathways (*e.g.* KEGG, Reactome), drug MOAs or Pfam

domains. For this, the ranked drug sets are converted into target gene/protein sets to perform Target Set Enrichment Analysis (TSEA) based on a chosen annotation system. Alternatively, the functional annotation categories of the targets can be assigned to the drugs directly to perform Drug Set Enrichment Analysis (DSEA). Although TSEA and DSEA are related, their enrichment results can be distinct. This is mainly due to duplicated targets present in the test sets of the TSEA methods, whereas the drugs in the test sets of DSEA are usually unique. Additional reasons include differences in the universe sizes used for TSEA and DSEA.

Importantly, duplications in the test sets of the TSEA are commonly caused by several distinct drugs sharing the same target proteins. Standard enrichment methods, such as those used for gene set enrichment, would eliminate these duplications since they assume uniqueness in the test sets. Removing duplications in TSEA would be inappropriate since it would erase one of the most important pieces of information of this approach. To solve this problem, I developed and implemented in the TSEA methods of *signatureSearch* a weighting method for duplicated targets, where the weighting is proportional to the frequency of the targets in the test set.

To perform TSEA and DSEA, drug-target annotations are essential. In *signatureSearch* they have been assembled from several sources, including DrugBank, ChEMBL, STITCH, and the Touchstone dataset from the LINCS project [215, 61, 104, 186]. Most drug-target annotations provide UniProt identifiers for the target proteins. If necessary, protein identifier sets can be mapped via their encoding genes to the chosen functional annotation categories, such as GO or KEGG. To minimize bias in TSEA or DSEA, often caused

by promiscuous binders, it can be beneficial to remove drugs or targets that bind to large numbers of distinct proteins or drugs, respectively. To conduct TSEA and DSEA efficiently, *signatureSearch* and its helper package *signatureSearchData*, provide several convenience utilities along with drug-target lookup resources for automating the mapping from drug sets to target sets to functional categories (Table 2.2). To avoid additional duplications caused by many-to-one relationships among protein isoforms and their encoding genes, most FEA tests involving proteins in their test sets are performed on the gene level in *signatureSearch*. For this, the corresponding functions in *signatureSearch* will usually convert target protein sets into their encoding gene sets using identifier mapping resources from R/Bioconductor, such as the *org.Hs.eg.db* annotation package. Because of this as well as simplicity, the following text and the corresponding documentation of the software will refer to the targets of drugs almost interchangeably as proteins or genes, even though the former are usually the direct, and the latter only the indirect, targets of drugs, respectively.

The following introduces the functionalities in *signatureSearch* for performing TSEA on drug-based GESS results using as functional annotation systems GO and KEGG pathways. For this the enrichment tests can be performed with three widely used algorithms that have been modified in *signatureSearch* to take advantage of duplication information present in the test sets used for TSEA. The relevance of these target duplications is explained above. To account for multiple hypothesis testing situations, the FEA functions support seven p-value adjustment methods. The Benjamini & Hochberg (BH) method is usually set as the default adjustment. The latter is used for the FEA tests included in this article [12]. First, I developed the *Duplication Adjusted Hypergeometric Test* (dup_hyperG).

This test is based on the hypergeometric distribution, which determines whether a discovered gene set shows an enrichment in functional annotations that is more extreme than what is expected from random sampling from the same gene universe [42]. To maintain the duplication information in this test, the size of the test set and number of proteins belonging to an annotation category (*e.g.* GO term) are both adjusted by the frequency of the target proteins in the test set. Effectively, the approach removes the duplications, but maintains their frequency information in form of weighting values. Second, I developed the *Modified Gene Set Enrichment Analysis* (mGSEA). The original GSEA method calculates the degree to which annotation categories are enriched at the extremes of ranked gene lists. For this an enrichment score is computed with a running sum Kolmogorov-Smirnov statistic and then evaluating significance by comparing the results to a null distribution derived from random queries [187]. To perform GSEA with duplication support, I am introducing in *signatureSearch* a modified GSEA (mGSEA) method, where the frequency information of targets is preserved by a weighting approach. Third, I implemented the MeanAbs (mabs) method in *signatureSearch*. MeanAbs is a simple but effective method for performing gene set-based enrichment analysis [46]. It assesses the enrichment of genes in an annotation category simply by averaging their absolute values of a chosen statistics (*e.g.* \log_2 ratios or Z-scores). Subsequently, significance is evaluated by comparing the result to a null distributions derived from random permutations of queries. The following describes the TSEA algorithms in more detail.

A. Duplication Adjusted Hypergeometric Test (dup_hyperG)

The classical hypergeometric test assumes uniqueness in its gene/protein test sets. Its p-value is calculated according to

$$p = \sum_{k=x}^n \frac{\binom{D}{k} \binom{N-D}{n-k}}{\binom{N}{n}}. \quad (2.1)$$

In case of GO term enrichment analysis the individual variables in equation (2.1) are assigned the following values. N is the total number of genes/proteins contained in the entire annotation universe, D is the number of genes annotated at a specific GO node, n is the total number of genes in the test set, and x is the number of genes in the test set annotated at a specific GO node. To maintain the duplication information in the test set used for TSEA, the values of n and x in the above equation are the corresponding gene counts including duplications.

B. Modified Gene Set Enrichment Analysis (mGSEA)

The original GSEA method [187] uses predefined gene sets S s defined by a chosen functional annotation system, such as GO or KEGG categories. The goal is to determine whether the genes in S are randomly distributed throughout a ranked test gene list L (*e.g.* all genes ranked by LFC), or enriched at the top or bottom of L . This is expressed by an Enrichment Score (ES) reflecting the degree to which a set S is overrepresented at the extremes of L . For TSEA, the test set L is a target set T associated with the top ranking drugs in a GESS result obtained from a drug-based GES database. Frequently, the corresponding gene identifiers in T are not unique, because several drugs in a GESS result may share the same targets. To account for the characteristic nature of GESS results, it is of utmost

importance to maintain this duplication information as much as possible. To perform GSEA with duplication support, here referred to as *mGSEA*, the target set T is transformed to a score ranked target list L_{tar} of all targets included in the corresponding annotation system. For each target in T , its frequency is divided by the number of all targets in T (including duplications), which is the weight of that target. For targets present in the annotation system but absent in the target set T , their scores are set to 0. Thus, every target in the annotation system will be assigned a score. Subsequently, the target list will be sorted decreasingly to obtain L_{tar} . Importantly, the original GSEA method cannot be used for TSEA directly since zeros are very frequent in L_{tar} . As a result, the sum N_R can become zero too which cannot be used as the denominator in equation (2.2) from Subramanian et al. (2005). To avoid this problem, the affected *ES* values are ignored by assigning -1 as a tag.

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p \quad (2.2)$$

If only some genes in set S have scores of zeros then the value of N_R is increased according to equation (2.3). The latter adds to N_R the minimum value of the non-zero gene scores in S multiplied by the number of genes in S that have scores of zero. Increasing N_R can in return decrease the weight of the genes in S that have non-zero scores. To compensate for this, the *mGSEA* algorithm computes N_R according to equation (2.3) instead of equation (2.2). $P_{\text{hit}}(S, i)$ in equation (2.2) evaluates the fraction of genes in S ("hits") weighted by their scores present up to a given position i in L_{tar} , where r_j is the score of gene j in L_{tar} . Typically, the exponent p is set to 1 in order to weight the genes in S by their scores in L_{tar} .

$$N_R = \sum_{g_j \in S} |r_j|^p + \min(r_j | r_j > 0) * \sum_{g_j \in S} I_{r_j=0} \quad (2.3)$$

The motivation for the above modifications is that if only a small number of genes in set S has non-zero scores and these genes rank high in L_{tar} , the weight of these genes will be close to 1 resulting in an $ES(S)$ of close to 1. Thus, the original GSEA method would score the gene set S of a functional category as significantly enriched. However, this is undesirable because in this example only a small number of genes is shared among the test target set T and the gene set S of a functional category. To avoid this, small weights are assigned to genes in S that have scores of zero. The latter decreases the weight of the genes in S that have scores other than zero, thereby decreasing the false positive rate. Finally, the functional categories (gene sets S s) are ranked by ES from highest to lowest, where the top ranking ones are favored as enriched GO terms and KEGG pathways.

C. MeanAbs (*mabs*)

The input for the *MeanAbs* method is L_{tar} , the same as for *mGSEA*. In this enrichment statistic, $mabs(S)$, of a gene set S is calculated as mean absolute scores of the genes in S [49]. In order to adjust for size variations in gene set S , random permutations (*e.g.* $\pi = 1000$) of L_{tar} are performed to determine $mabs(S, \pi)$. Next, $mabs(S)$ is normalized by subtracting the median of the $mabs(S, \pi)$ and then dividing by the standard deviation of $mabs(S, \pi)$ yielding the normalized scores $Nmabs(S)$. Subsequently, the portion of $mabs(S, \pi)$ that is greater than $mabs(S)$ is used as nominal p-value. Finally, the resulting nominal p-values are adjusted for multiple hypothesis testing using the Benjamini-Hochberg method [12].

Instead of translating ranked lists of drugs into target sets, as for TSEA, the functional annotation categories of the targets can be assigned to the drugs directly to perform Drug Set Enrichment Analysis (DSEA) instead. Since the drug lists from GESS results are usually unique, this strategy overcomes the duplication problem of the TSEA approach. This way the above described enrichment methods, such as GSEA or tests based on the hypergeometric distribution, can be readily accommodated in the underlying statistical methods without major modifications. As explained above, TSEA and DSEA performed with the same enrichment statistics are not expected to generate identical results. Rather, they often complement each other's strengths and weaknesses.

DTN Visualization

After identifying in drug-based GESS results enriched target classes via the above described FEA methods, it is important to visualize the results in graphical representations that are designed to simplify the functional interpretation of the analysis outcomes. To address this important need, *signatureSearch* provides functions to render the final results in form of interactive drug-target network representations.

In addition to network graphics, the *signatureSearch* package provides several other visualization and plotting functionalities. This includes visual summaries of GESS ranking scores (Table 2.2) which can be applied to selected perturbation types in GESS results across cell types and cell type classifications, such as normal and tumor cells. In addition, various visualization functionalities for FEA results are available, such as dotplots and gene-concept networks. To maximize shareability and extendability across open-source environments, visualization resources from other packages are integrated such as *clusterProfiler* [229].

2.2.8 Software Design

Integrating analysis software for GESS and FEA applications into an R/Bioconductor package has several advantages. First, Bioconductor provides access to a large number of high-throughput genome analysis tools that are interoperable by sharing the same data structures and S4 classes optimized for statistical analysis. Second, the approach simplifies the development of automated end-to-end workflows for conducting GESSs for many application areas. Third, it consolidates an expandable number of GESS and FEA algorithms into a single environment that allows users to choose the most appropriate methods and parameter settings for a given research question. Fourth, the usage of generic data objects and classes improves maintainability and reproducibility of the provided functionalities, while the integration with the existing R/Bioconductor ecosystem, such as the widely used `summarizedExperiment` class infrastructure, maximizes their extensibility and reusability for other data analysis applications. Fifth, it provides access to several community perturbation reference databases along with options to build custom databases with support for most common gene expression profiling technologies (*e.g.* microarrays and RNA-Seq).

Figure 2.2 illustrates the design of the package with respect to its data containers and methods used by the individual GES analysis workflow steps. Briefly, expression profiles from genome-wide gene expression profiling technologies (*e.g.* RNA-Seq or microarrays) are used to build a reference database stored in the Hierarchical Data Format 5 (HDF5). HDF5 is a technology that enables storage and efficient retrieval of very large data sets. For convenience the *signatureSearchData* package provides pre-built HDF5 reference databases for users. A search with a query signature against a reference database is initialized by

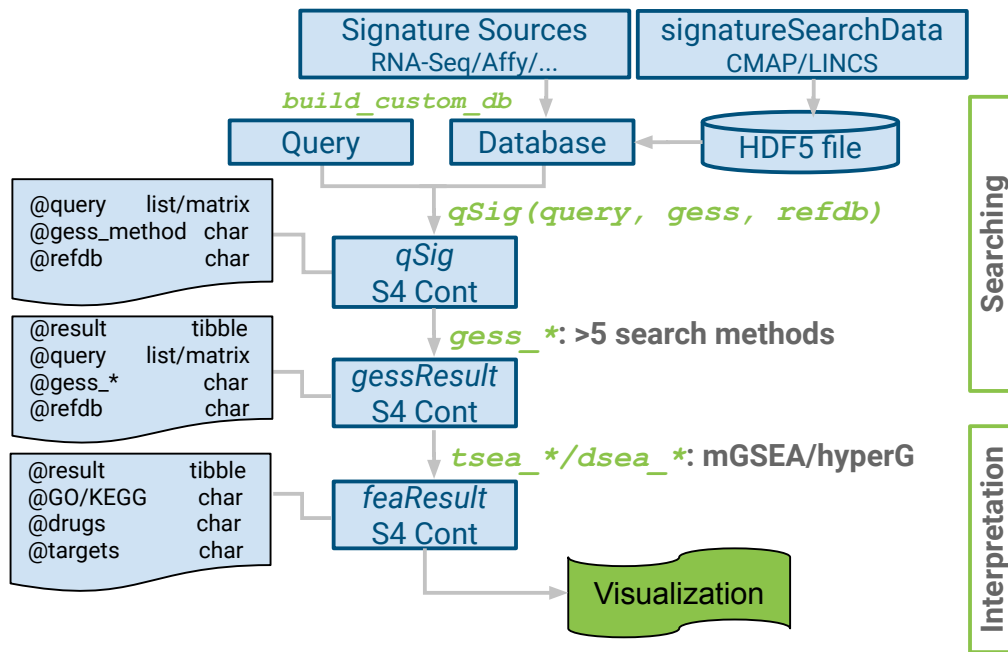


Figure 2.2: Design of *signatureSearch* package. GES reference databases are constructed from expression profile collections (RNA-Seq, Affymetrix chip or other technologies) and stored as HDF5 files. To perform GESSs, all query parameters are defined in a *qSig* search object where users can choose among over five search algorithms. The results are stored in a *gessResult* object that can be functionally annotated with different TSEA and DSEA methods. The enrichment results are organized in an *feaResult* object that can be used for drug-target network analysis and visualization.

Table 2.2: List of important functionalities provided by *signatureSearch* and *signatureSearchData*. The names of functions and libraries are italicized. ^aOnly the most common input types are listed. Acronyms are defined in the text.

Function Name	Description	Input ^a	Output and Comments
<i>(I) GES Databases</i>			
<i>CMAP2</i>	Affymetrix drug signatures	Raw, normalized and rank-based expression data	GES reference DB stored as HDF5 that can be accessed via <i>signatureSearchData</i> from <i>ExperimentHub</i> or user system
<i>LINCS</i>	L1000 drug & genetic signatures	Normalized and weighted averaged expression data	
<i>Custom</i>	User provided signatures	Many types of expression data	
<i>(II) GESS Methods</i>			
<i>gess_cmap</i>	CMAP method [109]	GS-Q: DEG; GEP-DB: Z-score/LFC ranks	<i>gessResult</i> object containing search result table with similarity scores for each perturbagen GES in the reference database, the query signature itself, as well as details about the chosen search parameters
<i>gess_lincs</i>	LINCS method [186]	GS-Q: DEG; GEP-DB: Z-scores	
<i>gess_gcmap</i>	gCMAP method [162]	GEP-Q: Z-score/LFC ranks; GS-DB: DEG	
<i>gess_fisher</i>	Fisher's exact test [68]	GS-Q: DEG; GS-DB: DEG	
<i>gess_cor</i>	Correlation methods [56]	GEP-Q; GEP-DB: same genes and GEP type	
<i>(III) FEA Methods</i>			
<i>tsea_mGSEA</i>	Modified GSEA algorithm [187]	Score ranked target list	<i>feaResult</i> object containing statistical enrichment results, details about chosen functional annotation system, labels of drugs used for testing, as well as their corresponding target information
<i>tsea_dup_hyperG</i>	Duplication adjusted hyperG test [42]	Target set with duplication	
<i>tsea_mabs</i>	meanAbs method [46]	Score ranked target list	
<i>dsea_hyperG</i>	Hypergeometric test [42]	Drug set	
<i>dsea_GSEA</i>	GSEA algorithm [187]	Score ranked drug list	
<i>(IV) Visualization</i>			
<i>gess_res_vis</i>	GESS result visualization	<i>gessResult</i> object	Dot plot of drug similarity scores
<i>comp_fea_res</i>	FEA result comparison	List of <i>feaResult</i> from FEA methods	Dot plot comparing result consistency
<i>dmnetplot</i>	Drug-target networks	Drug set; pathway ID	Interactive network graph

declaring all parameter settings in a `qSig` search object. Currently, users can choose here among five different search algorithms implemented in *signatureSearch*, while additional algorithms will be added in the future. The five implemented algorithms are listed in Table 2.2 and described in the previous section of this article. To minimize memory requirements and improve time performance, large reference databases are searched by sequential or parallel processing of its GES entries in batches of user-definable size. The search results are stored in a `gessResult` object that contains all information required to be processed by the downstream functional enrichment analysis (FEA) methods such as drug set and target set enrichment analysis (TSEA and DSEA) methods. The resulting functional enrichment information is organized in a `feaResult` object that can be passed on to various drug-target network construction and visualization methods implemented in *signatureSearch*.

2.3 Results

2.3.1 Performance Comparisons of GESS Methods

A. Test Design

Compounds with similar functional or structural properties are expected to induce GESs that are more similar among each other than those induced by compounds with dissimilar properties [186]. Based on this proof-of-concept assumption, I aim to systematically compare the performance of the six GESS methods, currently implemented in *signature-Search*, in recovering both functional and structural categories using known MOA categories and structure similarity clusters (SSC), respectively. That is, I ask the question: do drugs with similar molecular effects or structural features cluster in GESS results according to the corresponding classifications?

The MOA annotations used for these tests were downloaded from the Touchstone database [186]. The downloaded MOA annotations include 276 MOA categories and drug-target annotations for 1,555 drugs. Since not all of the MOA categories are expected to perform equally well in the GESS performance tests, the MOAs were ranked by their recall rates, and 25% of the top performers (here 69 MOAs with 309 drugs) were used for testing. To avoid bias in the final MOA selection, the recall rates were calculated across all GESS methods. Examples of poor performing MOA categories include those enriched in drugs that bind to sets of unrelated target proteins, or drug targets positioned far downstream of transcriptional regulation processes. In both cases the drugs of the corresponding MOA categories are not expected to induce related gene expression changes. Thus, including these

problematic MOAs in the recall performance tests would unnecessarily degrade the overall performance of the GESS methods.

The SSC categories were generated with the binning clustering method of the *ChemmineR* package [23]. This clustering step used atom pairs for structure similarity comparisons and the Tanimoto coefficient as similarity metric. For assigning compounds to clusters, a Tanimoto coefficient of 0.6 was used as similarity cutoff. The latter was chosen because it often generates, in combination with the atom pair method, clusters of reasonable size with relatively low numbers of false negatives and positives [203, 4]. Since PC3 cells had the best screening coverage in the LINCS database, the 5,253 compounds participating in the corresponding assays were used to generate the SSCs. Subsequently, the SSCs were filtered the same way as the MOA categories above, meaning only 25% of the top performers (here 139 SSCs with 542 compounds) were used for testing. The more details on the filtering procedure are provided in the following paragraph.

The 276 MOA categories were downloaded from the Touchstone database. They were associated with a total of 1,555 compounds. Since not all MOA categories are expected to contain drugs that induce similar gene expression changes, MOA categories predominantly associated with dissimilar GESs were eliminated by a filtering process based on recall rates that were averaged across all six GESS methods. For this, the GESs associated with drugs belonging to a MOA category were searched iteratively against the LINCS database. For each query result, the rankings of the GESs belonging to the same MOA category as the query were recorded. The joined ranking results for all queries of a MOA were then summarized using the mean of the ranks, and the mean rank percentile was set as the recall

rate of a MOA for the corresponding GESS method. To make sure none of the six GESS methods had been given an unfair advantage in this selection process, the MOA level recall rates were combined by calculating the mean of the recall rates across all six GESS methods. The latter was used for the final ranking of the MOA categories. Subsequently, the top 25% ranking MOA categories were used for the GESS performance tests described in the main text of this article. The final set included a total of 69 MOA categories associated with 309 compounds. The filtering of the SSC categories was performed the same way as the filtering of the MOA categories.

The GESs induced by the drugs in each MOA and SSC category were queried with each of the six GESS algorithms against the LINCS database and their similarity scores recorded for the corresponding database entries (Figure 2.3A). The query GESs of each drug used for the four set-based methods (*CMAP*, *gCMAP*, *Fisher* and *LINCS*) and the two correlation-based methods (*SPsub* or *SPall*) were the GSs corresponding to the 150 most strongly up- and 150 most down-regulated DEGs, and the GEPs subsetted to the same GSs or those for all assayed genes, respectively. The cell type, treatment time point and concentration chosen for these experiments were PC3, 24h and 10 μ M, respectively. Subsequently, the performance among GESS methods was compared in the form of receiver operating characteristic (ROC) curves by evaluating the true positive rate (TPR) against the false positive rate (FPR) across the full range of similarity scores obtained for each GESS method [156]. ROCs were computed for each GESS method by calculating their cumulative TPRs and FPRs from a binary vector that was sorted by the similarity scores of the combined query results (Figure 2.3B-C). In each binary result component, drugs

Table 2.3: GESS methods applied to MOA categories. To assess whether the observed performance differences are statistically significant for all pair-wise comparisons, the bootstrap method was used for both the global AUC and pAUC metrics. The BH method was used for multiple testing correction [12]. The columns contain: GESS method 1^a; GESS method 2^b; P-value^c and adjusted P-value^d.

		AUC		pAUC (FPR 0.01)		pAUC (FPR 0.05)		pAUC (FPR 0.10)	
GESS1 ^a	GESS2 ^b	P-Value ^c	P-Adjust ^d	P-Value	P-Adjust	P-Value	P-Adjust	P-Value	P-Adjust
gCMAP	CMAP	6.2e-70	1.3e-69	2.7e-11	2.9e-11	4.7e-29	5.4e-29	1.3e-39	1.5e-39
gCMAP	Fisher	1.4e-104	3.5e-104	1.3e-56	2.7e-56	1.6e-96	4.0e-96	1.1e-124	3.2e-124
gCMAP	SPall	8.1e-217	6.1e-216	9.8e-44	1.6e-43	3.7e-72	6.2e-72	2.9e-89	5.4e-89
gCMAP	LINCS	3.0e-182	1.5e-181	1.6e-97	4.9e-97	1.6e-193	2.4e-192	2.0e-211	1.5e-210
gCMAP	SPsub	7.0e-236	1.0e-234	6.4e-145	9.6e-144	3.3e-181	2.5e-180	8.9e-218	1.3e-216
CMAP	Fisher	1.7e-27	2.1e-27	1.1e-48	2.1e-48	1.0e-60	1.5e-60	3.2e-69	4.7e-69
CMAP	SPall	5.3e-65	9.9e-65	5.4e-38	8.1e-38	2.0e-37	2.5e-37	7.3e-42	9.1e-42
CMAP	LINCS	1.5e-132	5.8e-132	4.0e-86	1.0e-85	1.1e-151	5.3e-151	2.7e-159	1.3e-158
CMAP	SPsub	2.0e-128	5.9e-128	2.2e-125	1.6e-124	1.7e-144	6.3e-144	2.4e-141	8.8e-141
Fisher	SPall	1.2e-04	1.3e-04	1.5e-07	1.5e-07	9.1e-19	9.8e-19	1.4e-15	1.5e-15
Fisher	LINCS	2.8e-28	3.8e-28	3.9e-13	4.5e-13	7.8e-60	1.1e-59	4.4e-62	6.1e-62
Fisher	SPsub	2.3e-62	3.9e-62	2.3e-99	1.1e-98	1.2e-84	2.2e-84	7.6e-75	1.3e-74
SPall	LINCS	1.4e-16	1.6e-16	2.8e-22	3.4e-22	1.1e-85	2.4e-85	1.1e-90	2.3e-90
SPall	SPsub	1.1e-49	1.6e-49	9.1e-99	3.4e-98	1.1e-105	3.2e-105	1.3e-116	3.3e-116
LINCS	SPsub	1.0e-01	1.0e-01	5.2e-33	7.1e-33	5.7e-02	5.7e-02	1.8e-01	1.8e-01

from the same and different categories as the corresponding query were indicated with ones and zeros, respectively. The same ROC calculations were performed on MOA and SSC categories separately. In both cases this was done on both the category level and the global level by generating ROCs for each category separately and all of them combined, respectively. To quantitatively compare the ROC performance results, I calculated the Area Under the Curve (AUC) as well as partial AUCs (pAUCs). While the full AUC evaluates the performance over the entire range of GESS similarity scores, the pAUCs are used for testing early enrichment at specific FPRs, where I chose FPRs of 1%, 5% and 10%. To assess whether the observed performance differences are statistically significant for all pair-wise comparisons among AUCs and pAUCs (Figure 2.3D-E), the bootstrap method from Robin *et al.* [74, 156] was used combined with the Benjamini-Hochberg (BH) method for multiple testing correction [12]. The results of these tests are provided in Table 2.3 and 2.4.

Table 2.4: GESS methods applied to SSC categories. The column titles and content of this table are organized the same way as in Table 2.3.

		AUC		pAUC (FPR 0.01)		pAUC (FPR 0.05)		pAUC (FPR 0.10)	
GESS1	GESS2	P-Value	P-Adjust	P-Value	P-Adjust	P-Value	P-Adjust	P-Value	P-Adjust
gCMAP	CMAP	7.7e-183	1.9e-182	1.3e-10	1.4e-10	8.7e-28	1.0e-27	5.6e-34	6.0e-34
gCMAP	Fisher	2.0e-155	3.8e-155	8.1e-64	1.3e-63	9.6e-147	1.6e-146	1.6e-197	3.0e-197
gCMAP	SPall	0.0e+00	0.0e+00	1.6e-59	2.4e-59	3.0e-96	4.0e-96	2.2e-130	3.0e-130
gCMAP	LINCS	0.0e+00	0.0e+00	2.7e-173	8.0e-173	0.0e+00	0.0e+00	0.0e+00	0.0e+00
gCMAP	SPsub	0.0e+00	0.0e+00	4.5e-236	3.4e-235	0.0e+00	0.0e+00	0.0e+00	0.0e+00
CMAP	Fisher	3.7e-28	4.3e-28	3.2e-54	4.4e-54	1.1e-114	1.6e-114	2.2e-135	3.2e-135
CMAP	SPall	2.3e-84	3.2e-84	3.0e-52	3.8e-52	5.3e-64	6.7e-64	5.3e-78	6.6e-78
CMAP	LINCS	5.9e-190	1.8e-189	7.2e-166	1.8e-165	0.0e+00	0.0e+00	0.0e+00	0.0e+00
CMAP	SPsub	1.1e-275	4.0e-275	2.7e-251	4.1e-250	0.0e+00	0.0e+00	0.0e+00	0.0e+00
Fisher	SPall	1.3e-03	1.3e-03	8.2e-02	8.2e-02	6.9e-26	7.4e-26	3.8e-36	4.3e-36
Fisher	LINCS	4.6e-86	7.0e-86	8.0e-71	1.5e-70	7.2e-157	1.4e-156	1.0e-162	1.7e-162
Fisher	SPsub	7.7e-158	1.6e-157	7.3e-209	3.6e-208	1.1e-231	2.8e-231	1.1e-235	2.6e-235
SPall	LINCS	2.8e-53	3.5e-53	4.3e-72	9.3e-72	1.7e-185	3.7e-185	1.8e-200	3.9e-200
SPall	SPsub	2.5e-153	4.2e-153	5.1e-189	1.9e-188	8.2e-253	2.5e-252	1.4e-300	4.3e-300
LINCS	SPsub	3.8e-15	4.1e-15	3.1e-32	3.6e-32	7.9e-11	7.9e-11	1.0e-04	1.0e-04

B. Test Results

The distributions of the category level AUCs and pAUCs for MOAs and SSCs are shown in Figures 2.4A-B and C-D, respectively, in the form of violin plots that are sorted by the corresponding global AUC and pAUC values. Figure 2.4E summarizes the performance test results for MOA and SSC categories in form of ranks of AUC and averaged pAUC outcomes. The sums of the ranks (here height of stacked bars) reflect the final performance ranking of each GESS method.

According to the performance results in Figure 2.4, *SPsub* consistently shows the best recall performance for MOA and SSC categories with respect to both global and early enrichment. *LINCS* performs second best for the same performance metrics. The performance rankings of the other four GESS methods are also relatively consistent across the four AUC/pAUC metrics. Their final rankings in decreasing order are: *SPall*, *Fischer*, *CMAP* and *gCMAP* (2.4E). The corresponding bootstrap test results in Tables 2.3 and 2.4 indicate

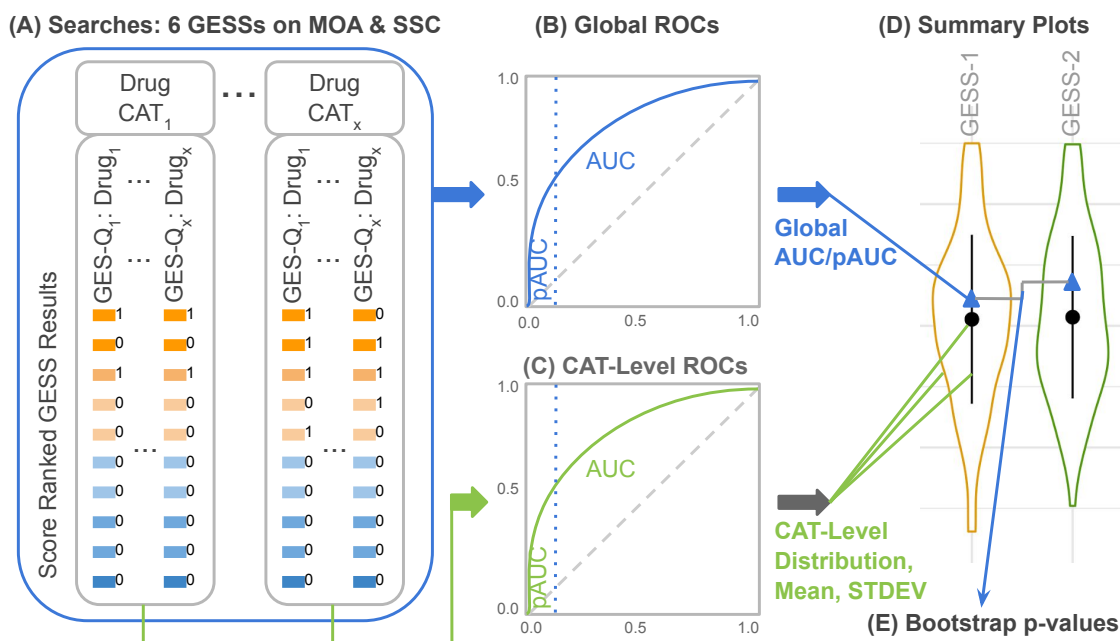


Figure 2.3: Performance testing strategy of GESS methods. (A) The GESSs of the drugs in each MOA and SSC category were searched against the LINCS database with each of the six GESS methods. The results were sorted by the corresponding similarity scores, here indicated by boxes with color gradient. GESSs from the same and different MOA/SSC categories (CAT) as the query were indicated in a binary vector with ones and zeros (next to boxes), respectively. After joining the binary vectors for each category group and re-sorting them by the corresponding scores, cumulative TPRs and FPRs were plotted in form of ROCs. This was done on the global level (B) and the CAT level (C) for the MOA and SSC classifications separately. (D) The distributions of AUC/pAUC values from each CAT-level are depicted by violin plots with mean values and standard deviation (STDEV) bars given in the middle. In addition, the global AUC/pAUC values are indicated by triangles. (E) The statistical significance of the observed differences among the global AUC/pAUC values of the six GESS methods was assessed by a bootstrap test described in the text.

that the observed differences among the AUC and pAUC values are statistically significant for nearly all pair-wise comparisons.

Among the correlation-based methods, *SPsub* performs better than *SPall* with respect to the AUC and pAUC performance metrics. One reason for this trend may be a lower noise level in the expression profiles used for computing the correlation coefficients for *SPsub* than *SPall*. The GEPs used for the *SPsub* method are usually enriched in genes (here most up- and down-regulated DEGs) that are robustly expressed, whereas the full gene repertoire used by *SPall* contains a larger proportion of genes with noisy expression signals. Among the set-based GESS methods, LINC performs best, while the classical Fisher’s exact test outperforms *CMAP* and *gCMAP* with respect to AUC and pAUC metrics for both MOA and SSC categories. The stronger performance of the *LINC* method compared to the other three set-based methods is most likely due to the additional weighting information utilized by this method.

Importantly, the global AUC values of the GESS methods are not expected to be very close to the best possible value of 1. However, they are in a high enough range to be substantially distinct from random assignments of drugs to MOA and SSC categories. In panel A and C of Figures 2.4, the global AUC values for MOA and SSC categories range from 0.53 - 0.72 and 0.54 - 0.77 with mean values of 0.65 and 0.68, respectively. It also has to be noted that the AUCs of the SSC categories are consistently higher than their MOA counterparts. This trend is expected because the SSCs were assembled with a single algorithm resulting in more homogeneous compound categories than the more complex annotation-based MOA classification system.

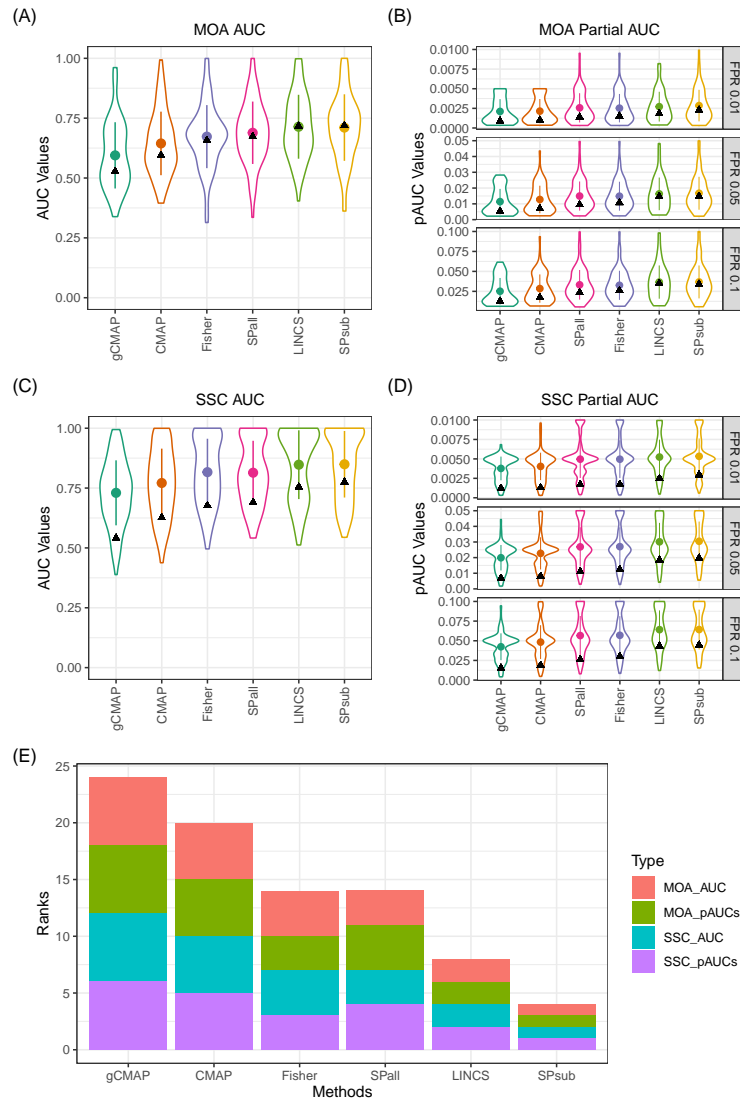


Figure 2.4: Recall performance of GESS methods on MOA and SSC categories. (A) The distributions of the ROC performance results of the 69 MOA categories are plotted in form of violin plots for each of the six GESS methods. The corresponding mean values, standard deviation bars and global AUCs are indicated within each violin by dots, vertical lines and triangles, respectively. The GESS methods are ordered by increasing global AUC values. (B) The corresponding distributions of pAUC values are given for FPRs of 1%, 5% and 10%. In this composite plot, the GESS methods are ordered by the mean of the ranks of their global pAUC values. (C)-(D) The GESS performance results of the 139 SSC categories are plotted the same way as the corresponding MOA results. (E) The performance results under (A)-(D) are summarized in form of stacked bar plots where the sum of the ranks is used to order the GESS methods from left to right by increasing performance. Each bar is composed of the ranking of the global AUCs and the mean ranking of the corresponding pAUCs for both MOA and SSC categories.

While the ranges of the AUC values for both classification types are reasonably high for real data, their absolute values should not be confused with a metric suitable for judging how well the majority of the GESS methods or the underlying GES assay technologies perform overall. Lower AUC values are expected for both category types mainly due to the complex nature of the real data set used for testing without degrading the reliability of the AUC-based ranking of the GESS methods. Clearly, the statistically significant ranking of the AUC values is the relevant information obtained from these tests. The following gives additional details why the maximum achievable AUC values are expected to be lower. First, the chosen MOA classifications are based on complex drug annotation data, which often do not have simple and unambiguous ground truth answers as it is possible with synthetic data. The structural similarity groupings of the SSC categories are also not expected to join compounds into groups, where every member is guaranteed to interact with the same targets or molecular processes. Second, the high noise level present in real large-scale mRNA expression data make correct assignments challenging, which in turn causes an additional reduction of the AUC values. Third, the presence of drugs binding to several targets from different MOA and SSC categories induces complex composite GESs. Finally, categories far up- or downstream of transcriptional control processes are unlikely to contain many drugs that recall each other to a high degree, no matter how well a GESS method performs overall. Despite these limitations, the MOA- and SSC-based GESS performance testing methods, chosen for this study, are appropriate choices in this use case, because they capture more biologically relevant information than alternative classification approaches based on synthetic data.

C. Time and Memory Performance

The GESS methods in *signatureSearch* process reference databases in batches with user-definable numbers of GES entries in each iteration of a full database scan. This allows searching of very large databases, while capping the memory consumption within the resources available on a computer system without major compromises on time performance. The time and memory performance of the six GESS methods is given in Table 2.5 for searching the LINCS database subsetted to ten thousand entries with a batch size limit of five thousand. The differences among the methods with respect to memory footprint and time performance for a fixed batch size is largely proportional to the size differences of the data required for each algorithm. For instance, the methods *SPsub* and *Fisher* only require for each GES entry the GSs of the most up- and down-regulated genes, whereas CMAP, LINCS and SPall require quantitative or rank-transformed GEPs for all assayed genes. Similarly, the processing times are shorter for the methods with more compact database entries, due to shorter load times when reading batches of GES entries into memory. The above time performance results are given for a single CPU core. If additional performance is needed (*e.g.* with very large databases), then it is easy to accelerate the search times by using the parallelization routines available in R/Bioconductor, such as *BiocParallel* or *batchtools* [14].

Table 2.5: Time and memory performance

GESS method	Time	Memory
CMAP	1.2min	3.5GB
LINCS	1.7min	2.3GB
gCMAP	1min	290MB
Fisher	9s	238MB
SPall	1min	838MB
SPsub	13s	238MB

D. Comparisons with Competing Software

This project implements commonly used GESS methods in a single environment including those that were previously only available as web services. Their performance has been compared above (Figures 2.3 and 2.4). Direct comparisons with web services are not an option for these tests, because they require large scale queries in the range of thousands of database searches with control over the GES database composition. Those requirements are usually not supportable by web services.

2.3.2 Use Case

The following demonstrates how the functionalities of *signatureSearch* can be applied to discovery-oriented research related to basic questions in biology, drug discovery and biomedical sciences. I selected as a query the GEP of SKB cells (skeletal muscle forming myoblasts) treated with vorinostat to search the LINCS expression database with the *SPsub* method. The latter GESS method was selected because it produced the strongest results in the above performance tests (Figures 2.3 and 2.4). Both the query (GEP-Q) and the entries in the reference database (GEP-DB) were based on pre-processed gene expression intensity values sub-setted to the 150 most up- and down-regulated genes from the vorinostat treatment of SKB cells. The drug vorinostat is a small molecule inhibitor of histone deacetylases (HDACs). Pharmacologically, it is used as antineoplastic agent and to treat cutaneous T-cell lymphomas (CTCL). It was chosen for this proof-of-concept test because several related HDAC inhibitor drugs with well annotated target annotations are represented in the LINCS database. Moreover, it has been used for similar reasons by other benchmark studies [109]

to determine whether GESSs of structurally and mechanistically related drugs are able to enrich each other at the top of GESS results.

Table 2.6 shows the top ten drugs of the vorinostat GESS result identified by *SPsub* and ranked by absolute correlation coefficients. Impressively, nearly all of the top ranking drugs are annotated to target the same or similar HDACs as the vorinostat query. Most importantly, the remaining two drugs in the table, KM-00927 and PCI-24781 (Abexinostat), are not yet annotated as HDAC inhibitors in the corresponding drug-target databases. However, two recent studies have identified them as novel HDAC inhibitors [172, 119]. PCI-24781 is an experimental drug candidate for cancer treatment, that has been approved for Phase II clinical trials for the treatment of B-cell lymphoma. It has also been identified as a novel hydroxamic acid-based HDAC inhibitor [155]. This result is an excellent example for demonstrating the power of the GESS technology in identifying targets for experimental drugs, as well as novel targets for drug repurposing approaches. Figure 2.5 compares the corresponding chemical structures of the compounds listed in Table 2.6. They are plotted in the order of a hierarchical clustering dendrogram generated with the structure-based clustering utilities of the affiliated *ChemmineR* package [23]. While it is not expected that GESS-based rankings will perfectly agree with structure-based rankings, at least in this case the compound groupings of the two methods are in reasonable agreement, as several compounds in Table 2.6 are indeed structurally related, such as PCI-24781, panobinostat, scriptaid and vorinostat.

Next, the top 100 drugs of the vorinostat GESS result were functionally annotated with the FEA methods implemented in *signatureSearch*. Since the results of the differ-

Table 2.6: Top ranking drugs of vorinostat query. The GES of SKB cells treated with vorinostat was used as query to search the LINCS database with the *SPsub* method. The rows are sorted decreasingly by absolute Spearman Correlation Coefficients^d. The other columns include ranks^a, drug names^b, cell types^c, and the gene symbols of the corresponding target sites^e.

Rank ^a	Drug Name ^b	Cell Type ^c	SCC ^d	Targets ^e
1	Vorinostat	SKB	1.00	HDAC1; HDAC10; HDAC11...
2	Trichostatin-a	SKB	0.99	HDAC1; HDAC10; HDAC2...
3	KM-00927	SKB	0.98	
4	Scriptaid	SKB	0.97	HDAC1; HDAC2; HDAC3...
5	HC-toxin	SKB	0.97	HDAC1
6	Belinostat	SKB	0.97	HDAC1; HDAC10; HDAC11...
7	Panobinostat	SKB	0.96	HDAC1; HDAC10; HDAC11...
8	PCI-24781	ASC	0.95	
9	HC-toxin	ASC	0.95	HDAC1
10	Vorinostat	ASC	0.94	HDAC1; HDAC10; HDAC11...

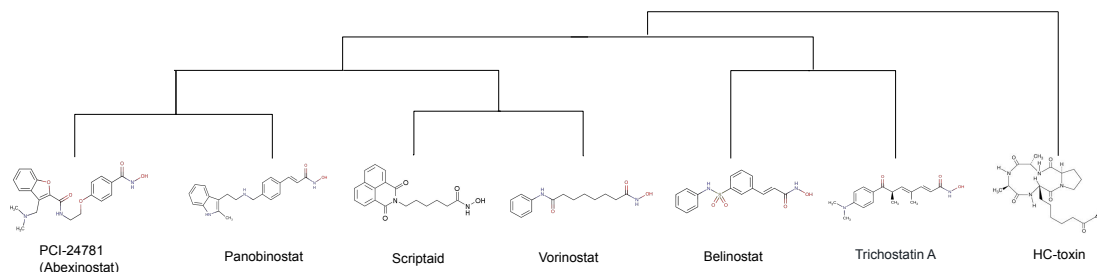


Figure 2.5: Structure-based hierarchical clustering dendrogram for drugs listed in Table 2.6. Experimental drugs lacking structure information are not included.

ent FEA methods were similar, the following considers only the results of the *dup_hyperG* method. Table 2.7 shows the 5 highest ranking GO terms of the Molecular Function (MF) and Biological Process (BP) ontology. The most highly enriched terms of the MF ontology are all related to histone deacetylase activity. This is expected since the target sites of the top ranking drugs are predominantly HDACs. The corresponding enrichment result for the BP ontology agrees well with the MF result since it is also dominated by histone deacetylation processes. Given vorinostat's HDAC inhibitor activity, the obtained FEA results demon-

strate the efficiency of *signatureSearch*'s FEA methods in identifying the correct pathways targeted by a query drug. Besides processes related to histone deacetylase activities, several biologically connected processes are enriched as well (Table 2.7), such as hair follicle placode formation. This is interesting because a recent study has shown that the suppression of epidermal HDAC activity leads to disrupted hair follicle regeneration and homeostasis [85]. This finding demonstrates the utility of the GESS/FEA workflow in identifying alternative target pathways that may enable novel drug repurposing approaches for query drugs of interest in the future. To highlight the importance of the FEA step in the overall workflow, Table 2.8 provides the enrichment results when using the genes of the initial GEP-Q instead of the downstream drug-target gene set from the GESS result for the same GO term enrichment analysis. When comparing the top ranking GO terms in both Tables 2.7 and 2.8 then there are no top ranking GO terms shared among the results. This is not surprising since the GES-Q contains the genes exhibiting the most pronounced expression changes after treating SKB cells with vorinostat, while the genes used for the FEA analysis are the genes encoding the target proteins of the top ranking drugs in the initial GESS search result. Typically, there are no or only minor overlaps expected among the genes in these two sets (here 1.6% of GEP-Q). Most importantly, only the FEA approach identifies the correct target pathway for the vorinostat query, whereas the GO term enrichment analysis with the genes from the initial GES-Q contains terms that are fundamentally different and unrelated to the vorinostat target pathway. This comparison demonstrates the critical role of the FEA method for the overall analysis workflow in predicting target pathways in drug-based GESS results with *signatureSearch*.

Table 2.7: Top ranking MF and BP terms obtained with *dup_hyperG*. The columns contain: GO Ontology^a; GO Term description/ID^b; number of proteins in GO term^c, test set^d and intersect^e, raw p-value^f, and adjusted p-value^g using the BH method for multiple testing correction. To save space, longer GO term descriptions have been shortened.

Ontology ^a	GO Term ^b	N GO ^c	N Test ^d	N Match ^e	P-Value ^f	P-Adjust ^g
MF	HDAC activity (H3-K14) (GO:0031078)	11	323	97	0.00e+00	0.00e+00
MF	NAD-dependent HDAC activity (H3-K14, GO:0032041)	11	323	97	0.00e+00	0.00e+00
MF	NAD-dependent HDAC activity (GO:0017136)	16	323	98	0.00e+00	0.00e+00
MF	NAD-dependent PDAC activity (GO:0034979)	17	323	99	0.00e+00	0.00e+00
MF	HDAC activity (GO:0004407)	44	323	98	0.00e+00	0.00e+00
BP	Histone H3 deacetylation (GO:0070932)	21	323	98	0.00e+00	0.00e+00
BP	Histone H4 deacetylation (GO:0070933)	11	323	59	0.00e+00	0.00e+00
BP	Histone deacetylation (GO:0016575)	86	323	101	0.00e+00	0.00e+00
BP	Hair follicle placode formation (GO:0060789)	5	323	23	0.00e+00	0.00e+00
BP	Fungiform papilla morphogenesis (GO:0061197)	5	323	23	0.00e+00	0.00e+00

Table 2.8: Top ranking GO MF and BP terms obtained from direct enrichment of the vorinostat GS-Q with hypergeometric test. The columns contain: GO Ontology^a; GO Term description/ID^b; number of genes in GO term^c, test set^d and intersect^e, respectively; as well as enrichment p-value^f and adjusted p-value^g using the Benjamini-Hochberg (BH) method. To save space, longer GO term descriptions have been shortened.

Ontology ^a	GO Term ^b	N GO ^c	N Test ^d	N Match ^e	P-Value ^f	P-Adjust ^g
MF	phospholipase activator activity (GO:0016004)	12	295	4	3.4e-05	0.013
MF	kinase regulator activity (GO:0019207)	207	295	13	4.6e-05	0.013
MF	lipase activator activity (GO:0060229)	14	295	4	6.6e-05	0.013
MF	transcription coactivator activity (GO:0003713)	319	295	16	9.6e-05	0.014
MF	RNA polymerase II TF binding (GO:0001085)	155	295	10	2.8e-04	0.024
BP	cellular response to peptide (GO:1901653)	385	293	21	8.3e-07	0.003
BP	regulation of apoptotic signaling pathway (GO:2001233)	406	293	20	7.1e-06	0.010
BP	histone modification (GO:0016570)	454	293	21	1.1e-05	0.010
BP	response to metal ion (GO:0010038)	364	293	18	1.9e-05	0.010
BP	covalent chromatin modification (GO:0016569)	474	293	21	2.1e-05	0.010

Subsequently, drug-target networks (DTNs) were constructed to visually interpret the FEA results and prioritize interesting candidate drugs. A sample DTN is shown in Figure 2.6 where the term ‘histone deacetylase activity’ (H3-K14 specific; GO:0031078) was chosen since it is one of the highest scoring GO MF terms in the previous result. The drugs and target proteins are depicted in Figure 2.6 as yellow boxes and circles, respectively, including vorinostat and its histone deacetylase targets. In the *signatureSearch* package these DTN graphs are fully interactive, where users can zoom into network modules, as well as select drugs and/or proteins in the drop-down menu located in the upper left corner of the plot.

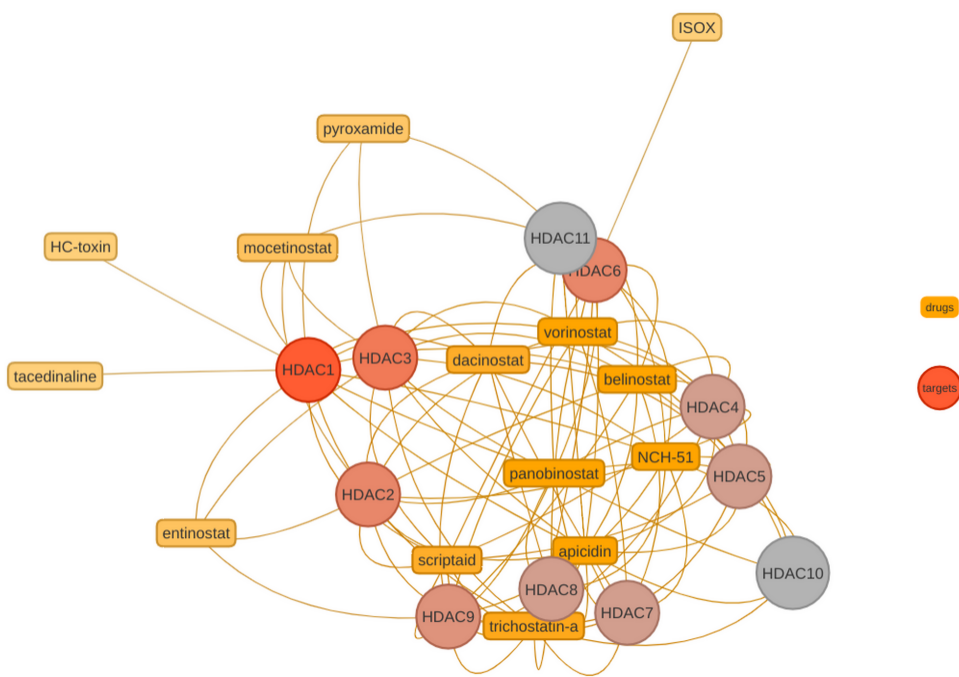


Figure 2.6: Drug-target network module of Histone Deacetylase Activity (H3-K14 specific; GO MF ID: GO:0031078). Drugs and targets are depicted as boxes and circles, respectively. The color of the circles indicates the number of connections.

2.4 Discussion

I have developed *signatureSearch* as an integrated and extendable environment for performing GESSs with a variety of algorithms combined with FEA and DTN visualization methods. The latter two are useful for guiding the downstream biological interpretation of GESS results. As outlined in the introduction and method sections, the software provides many useful and unique features, such as access to an end-to-end workflow toolkit covering most functionalities required for a wide range of GESS applications relevant to discovery-oriented research. It also provides access to an unmatched number of algorithms for both GESS and FEA routines, where I introduce several novel enrichment algorithms for interpreting GESS results. Importantly, the GESS methods in *signatureSearch* scale from single GES queries to large scale applications with thousands of GES queries using public or custom reference databases. This enables permutation tests with large numbers of randomized queries required to evaluate the robustness of GESS/FEA results. Typically, these types of large scale queries are not practical to support in other GESS tools that are predominantly based on web services. This study is also unique by testing the performance of the GESS algorithms in recalling MOA and SSC categories with drug-induced query GESs. To the best of our knowledge, the performance of GESS methods has not been systematically compared as it has been done here. In these performance tests I find that the correlation-based methods, *SPall* and *SPsub*, outperform most set-based methods with respect to the chosen ROC performance criteria. Among the set-based methods LINCS performs the best, most likely because of the additional weighting information utilized by its algorithm.

Although correlation-based GESS methods show the best performance in our tests, the query types required for them are more complex than the simple gene identifier sets required for the set-based methods. Moreover, for compatibility reasons the quantitative queries of correlation methods should preferentially be derived from the same gene expression technology and organism used for generating the reference database. In this regard the set-based methods are less restrictive and more versatile than correlation-based methods. Especially, for complex expression experiments, it is often easier to obtain a GES query composed of an identifier set of induced and repressed genes than the quantitative counterpart required for correlation-based approaches. Query gene sets from related species can also be used by translating them via ortholog mappings to the corresponding genes represented in the GES database. Moreover, set-based methods are more likely to exhibit reasonable performance in cross-omics queries, such as querying transcriptomic GES databases with up- and down-regulated gene sets from GWAS, proteomics or possibly even metabolomics studies. In summary, an advantage of set-based methods is that they are more technology agnostic but may not reach the recall performance of correlation-based methods. Integrating important GESS and FEA methods into an R/Bioconductor package also offers several unique advantages not present in related software applications. Here, the *signatureSearch* packages simplifies the development of automated end-to-end workflows for conducting signature searches in many application areas. It consolidates an extendable number of GESS and FEA algorithms into a single environment that allows users to compare results among methods as well as define and incorporate custom methods. Moreover, the usage of generic data objects and classes improves maintainability and reproducibility of the provided func-

tionalties, while the integration with the existing R/Bioconductor ecosystem maximizes their extensibility and reusability for other data analysis applications. Finally, *signatureSearch* provides access to several community perturbation reference databases along with options to build custom databases with support for most common mRNA expression profiling technologies. This design will also support expression profiling databases from other omics domains such as proteomics.

2.5 Conclusion

The *signatureSearch* package provides a general purpose environment for identifying similar GESs in reference databases, while also guiding the downstream functional interpretation of the discovered connections. The functionalities of the package pave the way for discovering biologically relevant connections in gene networks. Those are useful to gain insights into stress-response pathways, to improve treatments for diseases, or to identify novel target site candidates for experimental drug-like small molecules or alternative targets of approved drugs for drug-repurposing approaches. In the future I will continue to enhance the package by adding several new features. First, I will include additional GESS/FEA methods optimized and tested for sparse GES data, such as GESs from single cell sequencing experiments. Second, support will be added for managing large numbers of heterogeneous query GESs in a single container that can be populated from flat files or a custom query database. Third, a batch run function will be added to execute the GESS/FEA workflow on any number of these heterogeneous queries automatically. Fourth, support for community workflow environments, such as CWL and *systemPipeR* [72], will be

added to operate *signatureSearch* from start to finish from R or other popular programming languages such as Python or Bash.

2.6 Availability of Software and Data

signatureSearch and *signatureSearchData* are open source packages that have been reviewed, tested and accepted by the Bioconductor project. Both are freely available for all common operating systems from Bioconductor and GitHub: [signatureSearch](#) and [signatureSearchData](#).

Chapter 3

Application to Human Longevity

3.1 Abstract

Human longevity is influenced by genetic composition, environmental exposures, healthy diet and lifestyle choices. However, very little is known about longevity promoting biological processes that are accessible to pharmacological interventions. This study aims to reveal novel insights into longevity-associated drugs (LADs), genes (LAGs) and pathways (LAPs) by applying a combinatorial gene expression signature (GES) search strategy against the Integrated Network-based Cellular Signatures (LINCS) database. First, the performance of LINCS drugs, inducing GESs representative for their mechanism of action (MOA), was systematically assessed by computing for each MOA a recall score based on the GES similarity of the corresponding drugs. The obtained recall scores were used to prioritize LAD candidates in the downstream discovery steps. Second, longevity-associated MOA categories along with the corresponding drugs were identified by querying LINCS with GESs of drugs present in both the DrugAge and LINCS databases, and subsequently scoring the enrich-

ment of each MOA at the top of the ranked GES search results. The corresponding LAP candidates were identified via drug-target annotations available for each longevity-associated MOA category. Third, the most focused search results were generated by querying LINCS with the GESs from 11 well studied LADs as well as one longevity phenotype. To identify LAGs and LAPs, the target protein annotations of the newly identified candidate LADs were used for functional enrichment analysis. Finally, the results from the three steps were integrated and then interrogated with a combinatorial approach to select the most reliable set of novel LAD and LAP candidates. Collectively, this study identified a list of drugs, target proteins and pathways useful for pharmacological lifespan extension strategies.

3.2 Results

Figure 3.1 illustrates the analysis strategy in the human longevity research field. The GESS results from DrugAge LADs queries allowed to connect LADs with longevity associated MOAs. The GESs of well-characterized known LADs (‘Eleven LADs Selection’ method section) were searched against LINCS to identify novel LAD candidates triggering similar expression responses as the query LADs. FEA was performed on the top ranking LAD candidates to identify cellular networks and pathways that may be accessible to longevity-promoting pharmacological treatments. The resulting insights were used to identify novel LAD candidates useful for testing their effects on delaying or preventing aging related diseases in downstream assays. Taken together, this project identified putative LADs, LAGs and LAPs for drug development and repurposing to provide novel opportunities in the studies of pharmacologically modulating aging-related processes and diseases.

For readability and clarity, the following text uses acronyms for the drug sets used by this study. LAD87 refers to 87 DrugAge LADs that are also present in LINCS. This LAD87 set was used for connecting MOAs with longevity. LAD11 is a set of 11 high-confidence LADs used in the discovery and drug repurposing components of this study. The LAD11 drugs are also part of NIH’s Interventions Testing Program (ITP). They are listed in Table 3.1 along with their target, MOA, and other functional and therapeutic information. Finally, WCD3 refers to three well-characterized, non-longevity drugs (vorinostat, chlorpromazine and alvocidib) that were used for the proof-of-concept tests of the GESS/FEA workflow.

3.2.1 Recall Performance of MOA Categories

Dependent on the impact of the mechanisms of action (MOAs) of drugs on transcriptional processes, different MOAs are expected to exhibit variable performance in GESS applications. This is because drugs with MOAs directly perturbing transcriptional activities are more likely to induce GESs that are characteristic for each MOA than those largely disconnected from these processes. Moreover, drugs within the same MOA are expected to share more similar GESs than those from different MOAs. To systematically score the GESS performance of known MOAs, they were ranked by the recall performance of the corresponding GESs induced by drugs within each MOA (Figure 3.1A). The drug GESs were queried against LINCS database using the *SPsub* GESS method. The latter was chosen since it was the most accurate GESS method in performance tests conducted by Duan *et al.* (2020). At the time of this study, the LINCS database contained 334 and 138 MOAs each containing at least 2 or 5 drugs, respectively. Since recall performance tends to be robust for MOAs

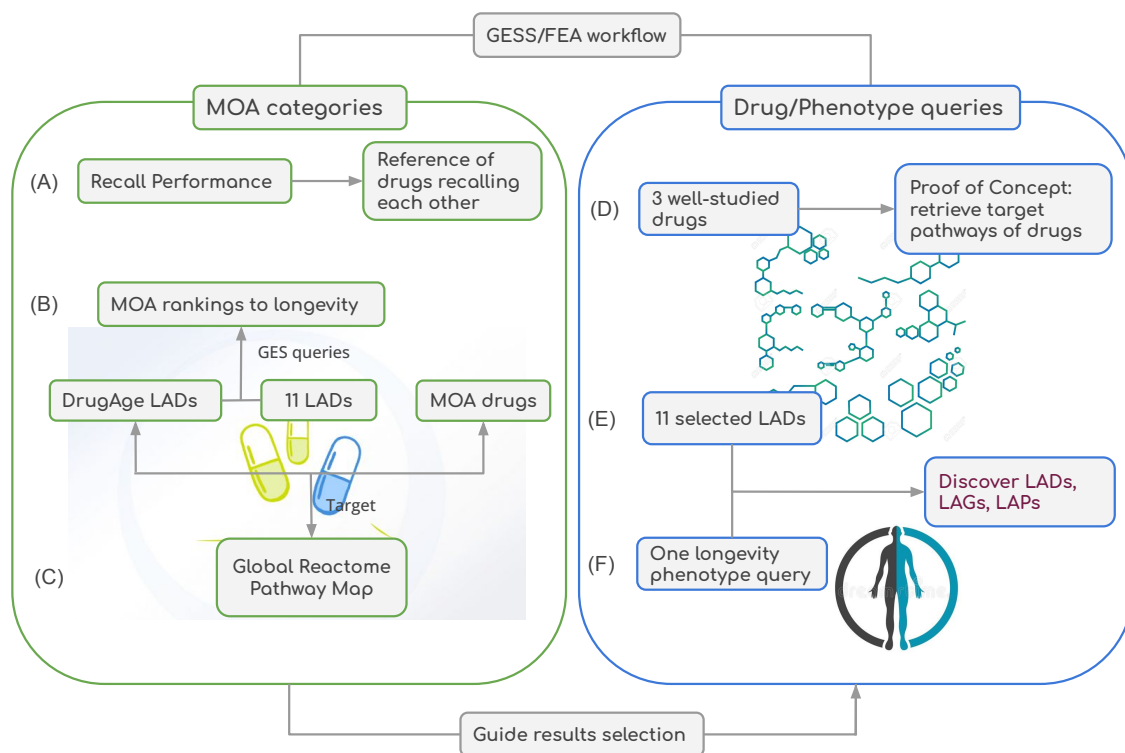













Figure 3.1: Strategy for identifying novel LADs, LAGs and LAPs. (A) The recall performance of MOA categories with at least 5 annotated drugs were estimated by searching the corresponding GESs against LINCS. The same MOAs were also ranked by their connectivity to longevity by using the GESs of known LADs as queries (B). The corresponding LADs included those present in both DrugAge and LINCS, as well as a custom set of 11 core LADs. Longevity associated target pathways were identified by enrichment analysis using Reactome where the target proteins of the above LAD sets were used as test sets (C). (D-F) An optimized GESS/FEA workflow was applied to three different GES sets from: (i) three well-characterized drugs as proof of concept experiment; (ii) 11 core LADs; and (iii) a longevity phenotype. The voting strategy was applied on 11 LADs GESs to get prioritized LADs candidates. The candidates can be flagged and associated with MOAs in the above recall performance tests.

Table 3.1: Overview of 11 high-confidence LADs. LAD data sources are: ITP¹: Interventions Testing Program from National Institute on Aging (NIA); Fuentealba *et al.* (2019)²; Strong *et al.* (2016)³. Additional annotation information is available in Table S1.

Drug	PubChem	MOA	Targets	N Targets	Structure
Sirolimus ¹ (Rapamycin)	5284616	MTOR inhibitor	CCR5, FGF2, FKBP1A, MTOR	4	
Tanespimycin ²	6505803	HSP90 inhibitor	HSP90A, HSP90B	2	
Acarbose ¹	41774	Glucosidase inhibitor	AMY2A; GAA; MGAM; SI	4	
Simvastatin ¹	54454	HMGCR inhibitor	CYP2C8; CYP3A4; CYP3A5; HMGCR; ITGB2	5	
Minocycline ¹	54675783	Inhibit protein synthesis	IL1B; ALOX5; VEGFA; CASP1; NOS2...	12	
Alpha-estradiol ³	68570	Estrogen receptor agonist	ESR1; ESR2; GPER1; NR1I2; CYP2B6...	11	
Beta-estradiol ¹	5757	Estrogen receptor agonist	ESR1; ESR2; GPER1; NR1I2; CHRNA4...	7	
Curcumin ¹	969516	Cyclooxygenase inhibitor; Histone acetyltransferase inhibitor; Lipoxygenase inhibitor; NFkB pathway inhibitor	APP; CA1; CA12; CA14; CA2; ...	21	
Resveratrol ¹	445154	Cytochrome P450 inhibitor; SIRT activator	SIRT1; APOA1; NQ02; CSNK2A1; PTGS1...	12	
Aspirin ¹	2244	Cyclooxygenase inhibitor	TP53; NFKBIA; EDNRA; AKR1C1; PTGS1...	19	
Metformin ¹	4091	Insulin sensitizer	ACACB; INS; PRKAB1	3	

with a minimum number of 5 drugs (Figure 3.9), the following uses MOAs with at least 5 drugs. In Figure S1 the 138 MOAs were ranked based on their recall rates, where lower values indicate better performance. The top ranking MOAs (Figure 3.2A) contain HDAC inhibitor, JNK inhibitor, nucleophosmin inhibitor, bacterial 30S ribosomal subunit inhibitor, MTOR inhibitor, tubulin inhibitor, inositol monophosphatase inhibitor, HMGCR inhibitor, DNA dependent protein kinase inhibitor, PARP inhibitor, *etc.* Many of these mechanisms are linked to transcriptional processes which in part explains their higher ranking as well as their lower dispersion of median CORct scores compared to lower ranking MOAs that more often act on processes far up- or down-stream from transcription.

Histone deacetylases (HDACs) are enzymes that remove acetyl groups from lysine residues of core histones, resulting in a more closed chromatin structure and repression of gene expression. HDAC inhibitors are a group of targeted anticancer agents that play important roles in epigenetic or non-epigenetic regulation, apoptosis, and cell cycle arrest in cancer cells [100]. C-Jun N-terminal kinase (JNK) signalling regulates both cancer cell apoptosis and survival. It is considered a potential oncogenic target for cancer therapy [219]. Nucleophosmin is a highly and ubiquitously expressed protein that plays crucial roles in ribosome maturation and export, centrosome duplication, cell cycle progression, histone assembly and response to a variety of stress stimuli. It is considered as a promising target for the treatment of both haematologic and solid malignancies [41]. Bacterial 30S ribosomal subunits defined an ‘assembly map’ for the order of association of each r-protein with rRNA. Their inhibitors are bactericidal antibiotics that act by binding to the subunit inhibiting bacterial protein synthesis, preventing tRNA attachment and also causing misreading of mRNA. Mechanistic

target of rapamycin (mTOR) is a protein kinase regulating cell growth, survival, metabolism, and immunity. It catalyzes the phosphorylation of multiple targets such as ribosomal protein S6 kinase β -1 (S6K1), Akt, protein kinase C (PKC), and type-I insulin-like growth factor receptor (IGF1R) to regulate protein synthesis, nutrients metabolism, growth factor signaling, cell growth, and migration [83]. Tubulin inhibitors act by a common mechanism via binding to the colchicine site on tubulin, which is a promising target for new chemotherapeutic agents for treatment of cancers, to inhibit tubulin assembly and suppress microtubule formation [120]. Inositol monophosphatase (IMPase) are involved in the phosphatidyl inositol (PI) signaling pathway, which affects a wide array of cell functions of cell growth, apoptosis, secretion, and information processing. HMG-CoA reductase (HMGCR) is the rate-limiting enzyme of cholesterol biosynthesis. HMGCR inhibitors (statins) are lipid-lowering medications clinically used to treat cardiovascular disease. The DNA-dependent protein kinase (DNA-PK) plays an instrumental role in the overall survival and proliferation of cells. As a member of the phosphatidylinositol 3-kinase-related kinase (PIKK) family, DNA-PK is best known as a mediator of the cellular response to DNA damage and an intriguing therapeutic target in the treatment of a variety of cancers. DNA-PK activity is also necessary for the regulation of transcription, progression of the cell cycle, and in the maintenance of telomeres [130]. Poly adenosine diphosphate-ribose polymerase (PARP) is a type of enzyme that helps repair DNA damage in cells. PARP inhibitors are a type of cancer drug in targeted therapy worked by preventing cancer cells from repairing, allowing them to die.

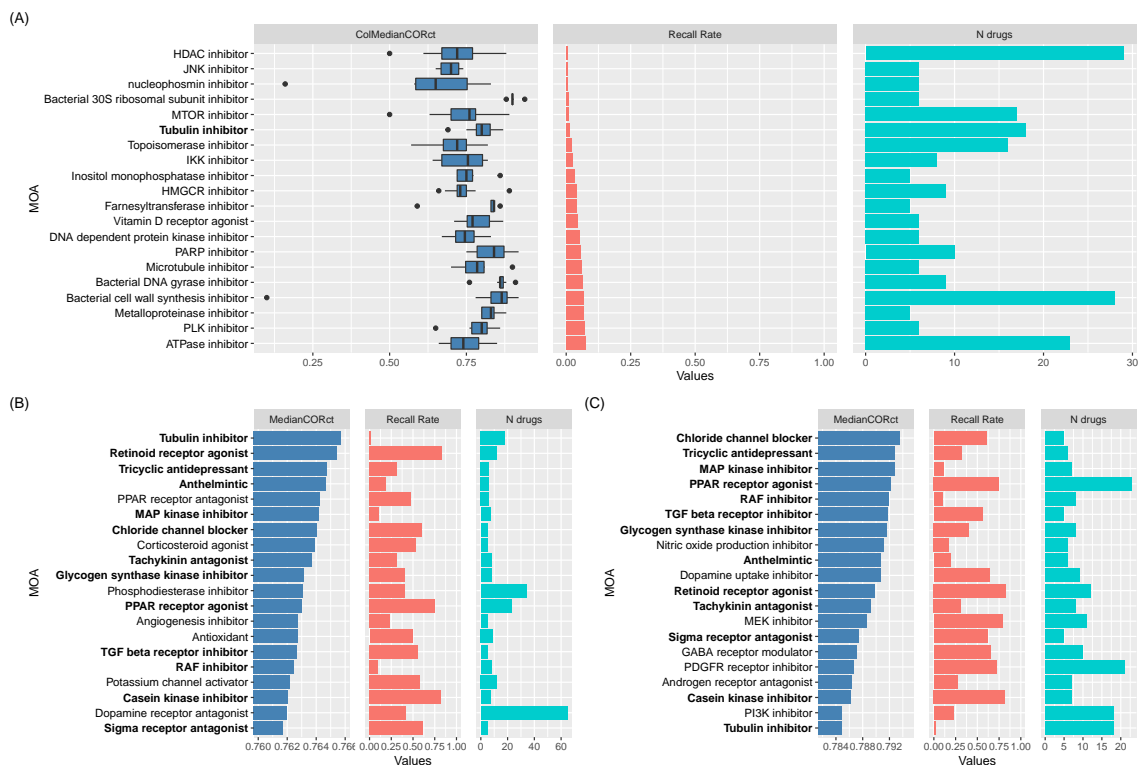


Figure 3.2: MOA recall rates and longevity association plots. (A) Top 20 MOAs ranked by recall performance. (B) Top 20 MOAs related to longevity ranked by median absolute CORct scores from GESs using LAD87. (C) Top 20 MOAs related to longevity from GESs using LAD11. Only MOAs with a minimum of 5 drugs are included for a total of 138 MOAs. Bold names indicate MOAs present in at least 2 panels. The complete MOA ranking results are available in Table S2, S3 and S4. ColMedianCORct values indicate the distribution of median absolute CORct for MOA drugs. Recall Rate are expressed as rank percentiles, N drugs represents the number of drugs in each MOA, and MedianCORct values indicate the median absolute CORct from LADs GES queries.

3.2.2 MOAs Connected with Longevity

Next, a comprehensive set of LAD-induced GESs was identified by intersecting drugs present in both DrugAge and LINCS (LAD87). DrugAge is a curated database containing small molecules with lifespan extending properties in a variety of animal systems [6]. The GEPs from LAD87 were used to query LINCS using the same GESS method as for the above recall performance tests. To score MOA categories by their connectivity to longevity (Figure 3.1B), the *CORct* scores of the search results were summarized to obtain a median score for each MOA that was used for ranking the MOAs. Since not all LADs reported in DrugAge may have longevity promoting effects in mammalian and human cells, a second LAD-based MOA ranking was generated using LAD11. The complete MOA ranking results are available in Table S2, S3 and S4.

The MOA to LAD connectivities are plotted for the top 20 ranking MOAs based on results obtained with both the LAD87 set (Figure 3.2B) and the LAD11 set (Figure 3.2C). The rankings of the top 20 MOAs are highly consistent among the two results (see bold labels in Figure 3.2B and C). The corresponding rankings of all MOAs in Table S3 and S4 are also similar as indicated by a pairwise Spearman correlation coefficient of 0.87. Because the recall- and LAD-based results were generated with query GESs of different drug combinations, their MOA rankings are not expected to agree among each other. In Figure 3.2A they are shown next to each other to assess whether high ranking MOAs in the LAD-based results are supported by strong recall performance. Many of the top ranking MOAs in the LAD connectivity results are functionally related to healthy aging processes and longevity. Examples include chloride channel blocker, tricyclic antidepressant, MAP kinase

inhibitor, PPAR receptor agonist, RAF inhibitor, TGF beta receptor inhibitor, glycogen synthase kinase inhibitor, anthelmintic, dopamine uptake inhibitor, retinoid receptor agonist, tachykinin antagonist, sigma receptor antagonist, casein kinase inhibitor, PI3K inhibitor, tubulin inhibitor.

Chloride channels are involved in a wide range of biological functions, including epithelial fluid secretion, cell-volume regulation, neuroexcitation, smooth-muscle contraction. Mutations in several chloride channels cause human diseases including cystic fibrosis, macular degeneration, kidney stones. Chloride-channel modulators have potential applications in the treatment of secretory diarrhoeas, polycystic kidney disease, osteoporosis and hypertension, some of these disorders are age-related [200]. So chloride channel might be a candidate target site to promote human healthy aging. Many studies have demonstrated that signaling pathways of MAPK, glycogen synthase kinase (GSK3, a key factor in growth and metabolism), PI3K, PPAR, RAF, TGF-beta (via insulin/IGF-1 signaling (IIS) pathway) can regulate longevity and delay age-associated metabolic disease [134, 179, 55, 221, 48, 176, 170], suggesting that they might be promising target sites in human to promote healthy aging and extend lifespan. Aging is accompanied with behavioral and cognitive decline. A study shows that serotonin and dopamine level decrease with age in *C. elegans* resulting in downregulation of the activity of neurons [226], dopaminergic neurons can regulate aging and longevity in flies [193]. Sigma receptors play a modulatory role in the activity of some ion channels and in several neurotransmitter systems, mainly in glutamatergic neurotransmission. Sigma receptor ligands have been proposed to be useful in several therapeutic fields such as amnesic and cognitive deficits, depression and anxi-

ety, schizophrenia, analgesia, and against some effects of drugs of abuse (such as cocaine and methamphetamine) [33]. Xu *et al.* (2019) found that the neuronal microtubule status might affect organismal aging through DAF-16-regulated changes in fat metabolism, and microtubule-based therapies might represent a novel intervention to promote healthy aging [220].

Moreover, it has been demonstrated that glucocorticoid signaling has contributed remarkably to therapeutic strategies in major organ systems in the human body [96] and drugs corresponding to this MOA (*e.g.* dexamethasone, diflorasone, fluocinolone, hydrocortisone, triamcinolone) are known to modulate pathways related to human longevity. In addition to these known aging related MOAs, a number of new longevity-associated MOAs are also identified including retinoid receptor agonist, and tricyclic antidepressant. Retinoic acid is important for developmental processes and cellular differentiation. It has antiproliferative and antioxidative properties, and regulates cellular differentiation [38, 168, 123]. Since the cellular effects of retinoic acid are mediated by the retinoid receptor, it is reasonable to hypothesize that the development of additional retinoid receptor agonists may lead to novel lifespan extending interventions. Moreover, tricyclic antidepressants may extend lifespan via directly impacting cellular function by activating non-cell autonomous stress responses or may alter neurophysiological responses that overall, may promote a healthy aging process [91, 152, 146]. However, the effect of some MOAs on promoting longevity and healthy aging may depend on the pharmacokinetics of the compounds that target these pathways. For example, while nitric oxide (NO) is an important antiinflammatory signaling molecule that mediates cellular function, high levels of NO bioavailability result in the gener-

ation of peroxynitrite and cellular damage [10, 135, 169]. Therefore, the development of NO production inhibitors may require dosage optimization. The effect of targeting longevity-associated MOAs may also depend on preexisting conditions or the route of administration. For example, corticosteroid agonists elicit antiinflammatory effects that extend lifespan and quality of life in individuals with emphysema yet increase risk of morbidity and mortality in individuals with rheumatoid arthritis [18, 30]. Regardless, the scope of the identified MOAs and their relation to longevity is broad and includes a variety of cellular pathways that are interesting targets for future longevity research.

3.2.3 LAD Targets within Global Pathway Map

To visualize and interpret the above MOA results, pathway enrichment analysis was performed and projected onto a global pathway map including the 28 highest level human pathways available in the Reactome database (Figure 3.1C, Figure 3.3). Pathways enriched in the target proteins of the LAD11 set are highlighted as colored branches in a fireworks plot (Figure 3.3A). To contrast enrichment differences among the three target sets (LAD11, DrugAge drugs and MOA drugs), their enrichment results were compared for the 28 high level as well as the full set of 2,508 descendant pathways. The former are shown as heatmap in Figure 3.3B, and the detailed results are available in Table S5. Descendant pathways enriched predominantly in the target set of the LAD11 (Figure 3.3A) include: SUMOylation (part of metabolism of proteins); activation of AMPK downstream of NMDARs (part of neuronal system); activation of PPARGC1A and mitochondrial biogenesis (part of organelle biogenesis and maintenance); MTOR signaling and extra-nuclear estrogen signaling (part of signal transduction); translocation of SLC2A4 (GLUT4) for glucose

transport (part of vesicle-mediated transport); and signaling by interleukins (part of immune system). The targets of LAD11 and DrugAge drugs are significantly globally mapped to gene expression (transcription) (descendant pathways are related to regulation of TP53 activity) and vesicle-mediated transport compared to a broader targets of MOA drugs, suggesting that the genes/proteins regulating gene transcription are the main target site for longevity-promoting drugs design, such as sirolimus is designed to target mTOR, which is a key regulator of gene transcription.

Several of the above mentioned processes are known be involved in aging and longevity in human and other organisms. Interestingly, SUMOylation is involved in cellular senescence and aging in human cell lines [167]. It has been shown in *C. elegans* to promote longevity and mitochondrial homeostasis [149]. AMPK controls the regulation of cellular homeostasis, metabolism, resistance to stress, cell survival and growth, cell death, autophagy, which are some of the most critical determinants of aging and lifespan. AMPK activation is shown to delay aging and prolong lifespan in *Drosophila melanogaster* [180]. AMPK activation in the *Drosophila*'s nervous system induces autophagy both in the brain and the intestinal epithelium, which is related to the anti-aging effects and extended lifespan [196]. Many reports demonstrate that AMPK activation and AMPK responsiveness decrease with age, which may explain the altered metabolic regulation, resulting in reduced autophagic clearance of unnecessary products (via mTOR), an increase in oxidative stress and decrease resistance to cellular stress (potentially due to DAF-16/FoxO and/or p53 signaling pathways downregulation). Thus, finding efficient strategies of increasing AMPK responsiveness and activation may be of important use as anti-aging treatments and for

lifespan elongation [161, 196, 20]. Mitochondria plays a key role in energy homeostasis and metabolism of reactive oxygen species (ROS), targeting pathways related to mitochondrial biogenesis might be of interest for extending human longevity [71]. Many studies have demonstrated that mTOR signaling have important influence on longevity and aging by influencing aging-related processes such as cellular senescence, cell growth, metabolism, and stem cell function [208, 138, 222, 92, 93, 83]. Extranuclear sex steroid receptors have been found in many normal cells and in epithelial tumors, where they enact signal transduction that impacts reproductive cycle, physiological stress responses, sleep cycle, and many other nonsexual behaviors [114, 67]. Glucose transporters (GLUTs) is involved in regulating tissue-specific glucose uptake and metabolism in the liver, skeletal muscle, and adipose tissue to ensure homeostatic control of blood glucose levels. Reduced glucose transport activity results in aberrant use of energy substrates and is associated with insulin resistance and type 2 diabetes. Studies establish that GLUT2 and GLUT4 are critical contributors in the control of whole-body glycemia [26]. Interleukins (ILs) are a group of cytokines modulating immune system. It is one of the main signaling pathways modulating the complex relationship between aging and chronic morbidity. The IL-6 pathway appears to be profoundly implicated in the pathophysiology of physical function decline and chronic diseases that often affect older persons. The modulation of IL-6 production or effects could offer a major breakthrough in prevention and treatment of people at advanced old age [122]. It may be possible to delay age-related diseases and aging itself by suppressing pro-inflammatory molecular mechanisms or improving the timely resolution of inflammation [153]. The transcription factor p53 plays a critical role in tumor suppression. In response to stress signals, p53 regulates its target

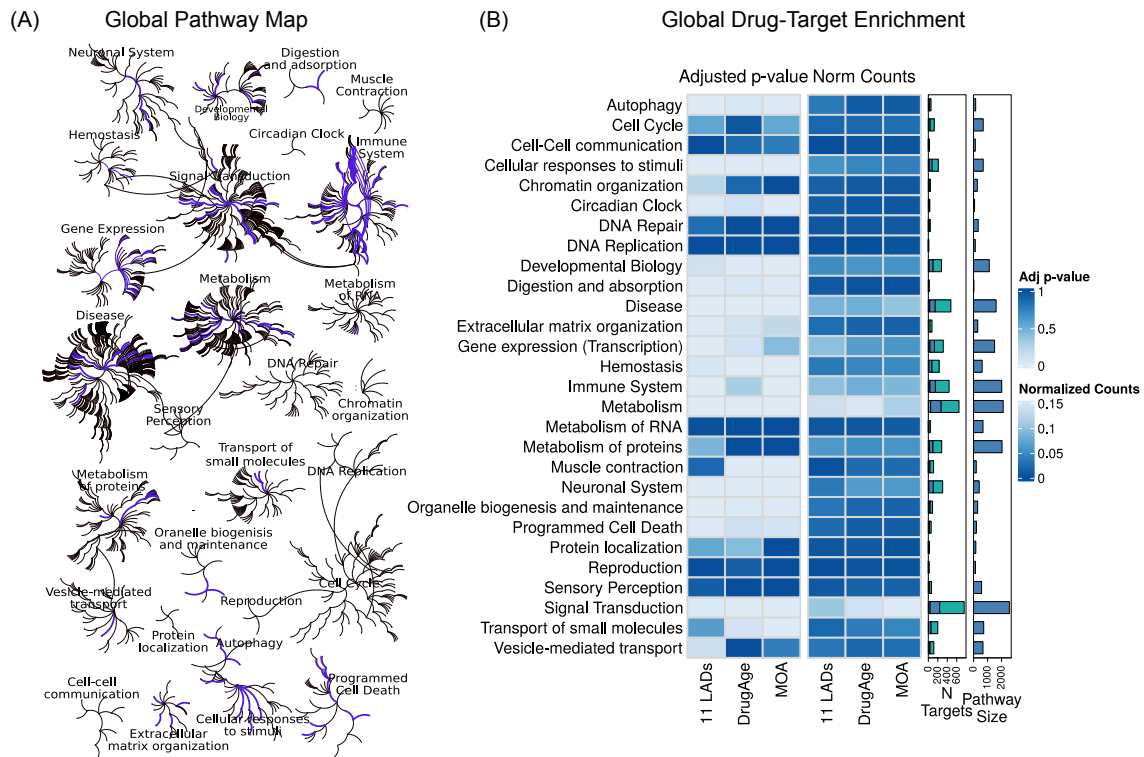


Figure 3.3: Global pathway mapping of drug targets. (A) Fireworks plot of the 28 highest level pathways in Reactome. Blue branches indicate enrichment p-values ≤ 0.01 for targets of the 11 LAD set. (B) Heatmaps of enrichment results for 28 high level pathways (rows) and 3 target sets (columns) are given for adjusted p-values of hypergeometric distribution test and normalized pathway mappings on left and right, respectively. Adjacent bar plots give the number of targets for each query set and the total number of genes in each pathway.

genes and initiates stress responses, including cell cycle arrest, apoptosis, and/or senescence, to exert its function in tumor suppression. Emerging evidence has suggested that p53 is also an important but complex player in the regulation of aging and longevity in worms, flies, mice, and humans in a context-dependent manner [52, 197, 194, 9, 199, 148, 2]. The potential mechanisms by which p53 regulates aging and longevity including the p53 regulation of IGF-1/AKT/mTOR signaling, stem/progenitor cells, and reactive oxygen species [51, 82].

3.2.4 Optimization of GESS/FEA Workflow

To discover novel candidate LADs and target LAPs, the GESSs of a diverse set of well characterized LADs were used to perform GESS against the LINCS database. Subsequently, functional enrichment analysis (FEA) was used to associate the resulting lists of the highest ranking drugs with target proteins and pathways. The combined analysis workflow is termed as GESS/FEA. To demonstrate the feasibility and efficiency of the approach (Figure 3.1D), the GESS/FEA workflow was first tested on WCD3 in SKB (muscle), SKB (muscle), and NPC (central nervous system) cell types respectively (Supplementary Sections). In this test the assignment of the correct drugs and target pathways in GESS and FEA results served as proof-of-concept. The results demonstrate that the *LINCS* GESS methods can identify compounds that share similar targets/MOAs as the query drugs and the utility of the chosen GESS/FEA workflow can discover the correct functional pathways of the query drugs.

3.2.5 GESS and FEA Results for Eleven LAD Signatures

GESS Results Summary

The GESS/FEA workflow was then applied to GESSs of a diverse set of LAD11 to discover novel candidate LADs and LAPs after validated with WCD3. In these analyses the robustness of the obtained GESS/FEA results was assessed via a rigorous permutation test computing Rank Robustness Scores (RRSs). Table 3.2 and Table 3.3 show the sample GESS and FEA results for one sirolimus GESS query. In order to understand the relationship between drugs that may translate into animal studies, the query GESSs for the LAD11 include both in vitro (GESSs drawn from LINCS database treated in human cell lines) and in vivo

(GESs from RNA-Seq or Microarray technologies treated in mice tissues) samples. It results a total of 56 query GESs for LAD11. Table 3.4 shows the sample information, the starred cell types are selected when the LAD names are used as sample names in the results.

To demonstrate a substantial difference from random, the in vitro GESS results from the LAD11 GESs in starred cell types were compared with random queries. As shown in Figure 3.4A and validated with WCD3 (Figure S2A), in general, the top ranking GESS results of the LAD11 have larger NCS scores as a measure of GES similarity than random GES queries. Next, the relationship between the GESS results and the identification of drugs sharing MOAs with the query drug was explored and validated with WCD3 (Figure S2B). In this analysis, GESS results were ranked by their absolute value of NCS scores and unique by the compounds. The rankings of drugs in the GESS results were then highlighted if they share a MOA with the query drug. As illustrated in Figure 3.4B, in general, drugs sharing MOAs with the query drugs tend to be enriched in the top ranking GESS results, demonstrating the ability of GESS method in retrieving compounds that share similar targets/MOAs as queries and suggesting that the similarity between compounds identified in GESS may relate to the similarity between query-GESS result MOAs. Next, the LAD11 were hierarchically clustered by their GESS results ranking similarity. Resolved clusters were then annotated with cluster specific overlapping reactome pathways (Figure 3.4C). These results indicated that acarbose, resveratrol, tanespimycin, minocycline in cluster 1 are involved in neutrophil degranulation (R-HSA-6798695), sirolimus, simvastatin and aspirin in cluster 2 are involved in signaling by interleukins (R-HSA-449147), interleukin-4 and interleukin-13 signaling (R-HSA-6785807) and G alpha (i) signalling events (R-HSA-418594),

and curcumin, metformin, alpha-estradiol, beta-estradiol in cluster 3 are involved in fatty acid metabolism (R-HSA-8978868), PIP3 activates AKT signaling (R-HSA-1257604), intracellular signaling by second messengers (R-HSA-9006925) and diseases of signal transduction by growth factor receptors and second messengers (R-HSA-5663202). The LAD11 were also clustered by other metrics including structural similarity, Jaccard index of target proteins and pathways (Supplementary Sections). Correlation analyses between GESS results from in vitro and in vivo query samples for sirolimus, estradiol and acarbose treatments were performed. The results indicate that the in vitro and in vivo samples have more correlation within them compared to correlation between in vitro and in vivo samples (Figure 3.4D).

In order to get prioritized drugs (PDs) in GESS results for each query LAD summarized across different cell types and technologies, several drug prioritization methods (DPMs) were proposed and tested (Supplementary Sections). The VoteNCSUnique method with drug classification was demonstrated as the most effective method in retrieving drugs sharing MOAs with the query LAD in their top rankings. The following shows the drug prioritization results for each of the LAD11 by using the VoteNCSUnique method, which computing a summary score for each LINCS drug ranking in the top N positions across many GESS results (*i.e.* cell types), and then re-ranking the results accordingly. Figure 3.4E, 3.4F, and S5-7 show the top 50 PDs for each of the LAD11. In addition to the 11 query LADs, many other famous drugs in DrugAge database are prioritized in the list including wortmannin (a fungal metabolite that was identified as a potent and selective inhibitor for phosphoinositide 3-kinases (PI3Ks) and PI3K-related enzymes), cinnarizine (antihistamine used for motion sickness), SU-4312 (selective and potent vascular endothelial cell growth factor re-

ceptor (VEGFR) inhibitor), arctigenin (a plant lignan with antioxidant, anti-inflammatory, anti-cancer and antiviral activities), dihydroergocristine (an ergot alkaloid that has a partial agonist activity on dopaminergic and alpha-adrenergic receptors, antagonist activity on serotonin receptors. The drug was approved by FDA for the treatment of dementia like Alzheimer's), lonidamine (a drug that interferes with energy metabolism of cancer cells, principally inhibiting aerobic glycolytic activity by its effect on mitochondrially-bound hexokinase (HK)), quercetin (a plant pigment that belongs to flavonoids), geldanamycin (an anti-tumor antibiotic that potently inhibits the function of Heat Shock Protein 90 (HSP90) that play important roles in the regulation of the cell cycle, cell growth, cell survival, apoptosis, angiogenesis and oncogenesis), CGP-52411 (a selective inhibitor of the epidermal growth factor receptor (EGFR), also inhibits and reverses the formation of A β 42 fibers associated with Alzheimer's disease, reduces neurotoxicity by blocking Ca²⁺ influx into neuronal cells), and staurosporine (an alkaloid isolated from *Streptomyces staurosporeus* exhibiting anti-cancer activity. It is a potent, non-selective inhibitor of protein kinases) that are classified into layer 1 (MOA match) or layer 2 (pathway match), they can be very promising drugs in extending lifespans for drug repurposing. Moreover, wortmannin, geldanamycin, lonidamine and dihydroergocristine are also prioritized from another summary approach across LAD11 (Supplementary Sections, Figure S8), SU-4312, staurosporine and quercetin are filtered from the drug-target (DT) network of signaling by VEGF Reactome pathway (Figure 3.6D), making them more convincing. Some of the drugs in layer 1 or layer 2 with therapeutic efficacy could also be interesting, such as AZD-8055, BMS-754807, amsacrine, itraconazole, cyclopentolate, leflunomide, clotrimazole, rosuvastatin, dextketoprofen, trogli-

tazone, estropipate, budesonide, mesalazine, fenbufen, diethylstilbestrol, tenoxicam, estrone, and gamma-linolenic-acid. Layer 3 contains a list of unknown drugs or small molecules that can be assumed as novel findings such as latrunculin-b, PI-828, MNITMT, U-0124, SA-1922006, MW-A1-12, oxetane, O-3M3FBS, SC-I-004, MW-SHH-61, isoflupredone-acetate, and THZ-2-98-01. The top 50 PDs for each of the LAD11 are then summarized into one heatmap by including drugs that are supported/prioritized by at least two LADs (Figure 3.5, Table S8). 12 drugs are supported by 3 query LADs including apigenin (a flavonoid found in many fruits and vegetables as well as in Chinese medicinal herbs, it has been widely investigated for its anti-cancer activities and low toxicity [223]), betulinic-acid (a natural pentacyclic triterpenoid with antiretroviral, antimalarial, and anti-inflammatory properties, has potential as an anticancer agent by inhibition of topoisomerase [31]), gamma-linolenic-acid (an omega-6 fatty acid, which the body can convert to substances that reduce inflammation and cell growth), hyperforin (a phytochemical that exhibits antidepressant activity, antibiotic activity against gram-positive bacteria, and antitumoral activity in vivo), mepacrine (an acridine derivative initially used for malaria and later as an antiprotozoal and immunomodulatory agent), mevastatin (a cholesterol-lowering agent), NU-7026 (an ATP-competitive inhibitor of DNA-dependent protein kinase (DNA-PK)), rosuvastatin (used along with a proper diet to help lower "bad" cholesterol and fats (such as LDL, triglycerides) and raise "good" cholesterol (HDL) in the blood), serdemetan (an orally bioavailable HDM2 antagonist with potential antineoplastic activity. It inhibits the binding of the HDM2 protein to the transcriptional activation domain of the tumor suppressor protein p53), sulindac (a nonsteroidal anti-inflammatory drug used to treat mild to moderate pain and help relieve

Table 3.2: Top 10 ranking drug cell terms in the GESS result from one sirolimus GES query in SKB cell with LINCS GESS method. RRS5: rank robustness score at 5 percent randomization.

Rank	Drug Name	Cell	WTCS	WTCS Pval	NCS	Tau	NCSct	Targets	RRS5
1	sirolimus	SKB	1.00	0.00e+00	2.80	99.96	1.22	CCR5; FGF2; FKBP1A; MTOR	1.00
2	BRD-K37940862	HT29	0.65	1.52e-05	1.97	98.38	1.19		1.00
3	BRD-K01614657	SKB	0.67	1.44e-05	1.87	97.73	1.03		0.80
4	SB-218078	SKB	0.66	1.46e-05	1.86	98.72	0.03	CHEK1	0.93
5	BRD-K82971429	SKB	0.66	1.46e-05	1.86	99.62	0.93		0.79
6	wortmannin	HT29	0.62	1.67e-05	1.86	99.19	1.33	ATM; ATR; MTOR; MYLK; PI4KA; ...	0.69
7	wortmannin	A549	0.60	1.73e-05	1.84	97.95	1.33	ATM; ATR; MTOR; MYLK; PI4KA; ...	0.66
8	amsacrine	HA1E	0.60	1.72e-05	1.83	98.84	1.22	ALB; KCNH2; ORM1; TOP2A; TOP2B	0.68
9	BRD-K16541732	SKB	0.65	1.52e-05	1.82	98.20	1.27		0.78
10	BRD-K61776140	SKB	0.65	1.53e-05	1.82	98.55	0.97		0.72

symptoms of arthritis), tenoxicam (an anti inflammatory analgesic used to treat mild to moderate pain) and WY-01-034 (PubChem CID: 70680403). Drug dihydroergocristine is supported by 2 query LADs of estradiol and simvastatin. Table 3.5 lists the annotations for all of the prioritized DrugAge drugs in this project, such as description, tested assays, and publications. Users could choose from these drug lists according there research interest that are worth testing in experiment studies. For example, the candidate drugs can be first tested in mouse fibroblasts and then in living mice for their efficacy in extending life span.

FEA Results Summary

In order to discover new pathways that are accessible to pharmacological life span extension strategies, I investigated which pathways are enriched among the molecular targets of the top ranking drugs in GESS results. This analysis was performed in several steps. First the ranking positions of the pathways directly targeted by the query LADs in their pathway rankings in FEA results were explored and validated with WCD3 (Figure S2C) to

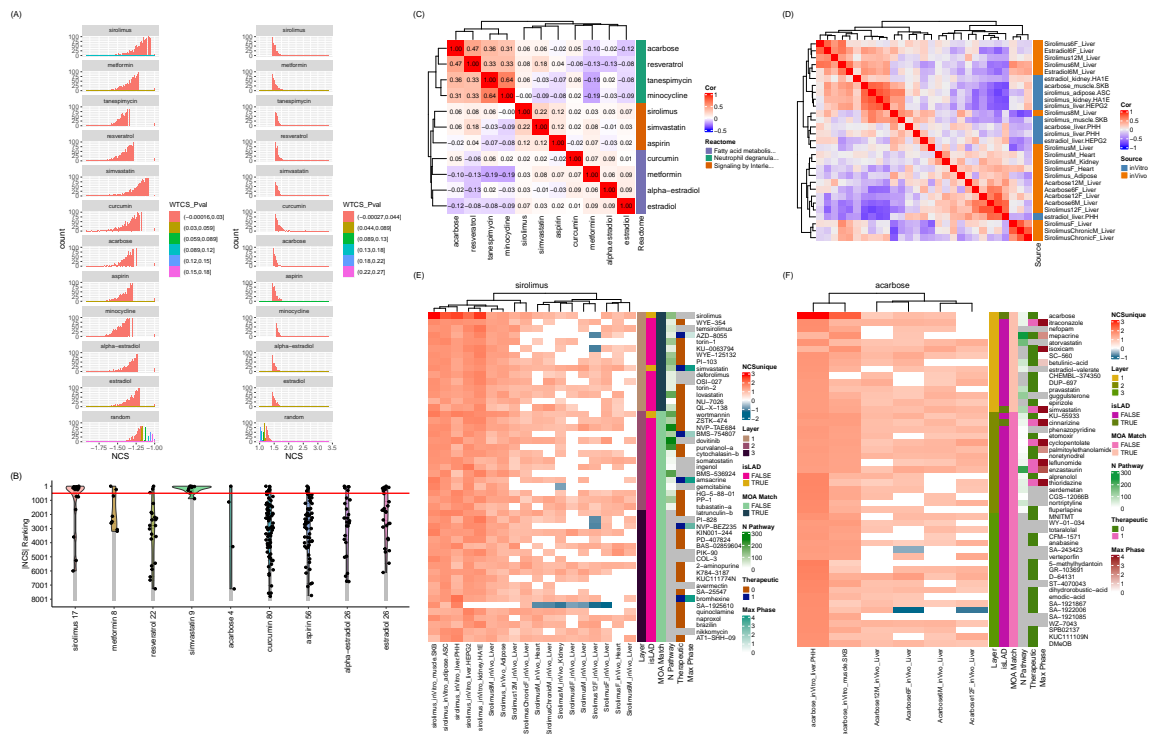


Figure 3.4: GESS result summaries across LAD11 queries. A: NCS score distributions in the GESS results from LAD11 queries and one random GES query as negative control for NCS scores less than -1.00 (left panel) and greater than 1.00 (right panel) after setting count cutoff as 100 to better show the NCS distributions at left and right extremes. The color key shows the P-values of the WTCS for the entries in the GESS results. B: GESS result rankings of drugs in the same MOA as the query LAD. Numerical values in the x-axis labels indicate the number of drugs sharing the MOA. Black dots represent drugs sharing the same MOA as query LAD and Grey bars indicate the total number of drugs in GESS results. C: hierarchical clustering of LAD11 by their GESS results ranking similarity. The color key indicates the Spearman correlation coefficient of the GESS result rankings from NCS scores after filtering of zeros. D: Clustering of in vitro and in vivo samples from sirolimus, acarbose and estradiol queries by their GESS results ranking similarity. The color key is the same as Figure C. E and F: Top 50 PDs from voteNCSunique method with stratification on LADs queries of sirolimus and acarbose, respectively. The columns are query samples that are clustered by Euclidian distance of NCSunique scores. The compound annotations plotted in the right bars include layer information (Layer), whether in DrugAge database (isLAD), whether the compound share at least one MOAs with MOAs of LAD11 (MOA Match), the number of compound targeted Reactome pathways shared with target Reactome pathways of LAD11 (N Pathway), whether the compound is therapeutic (Therapeutic), Max phase study by FDA (Max Phase). The complete tables containing all compounds rankings from voting strategy with scores, layer and annotation information corresponding to the heatmaps are stored in Synapse ([syn27074560](https://synapse.org/syn27074560)).

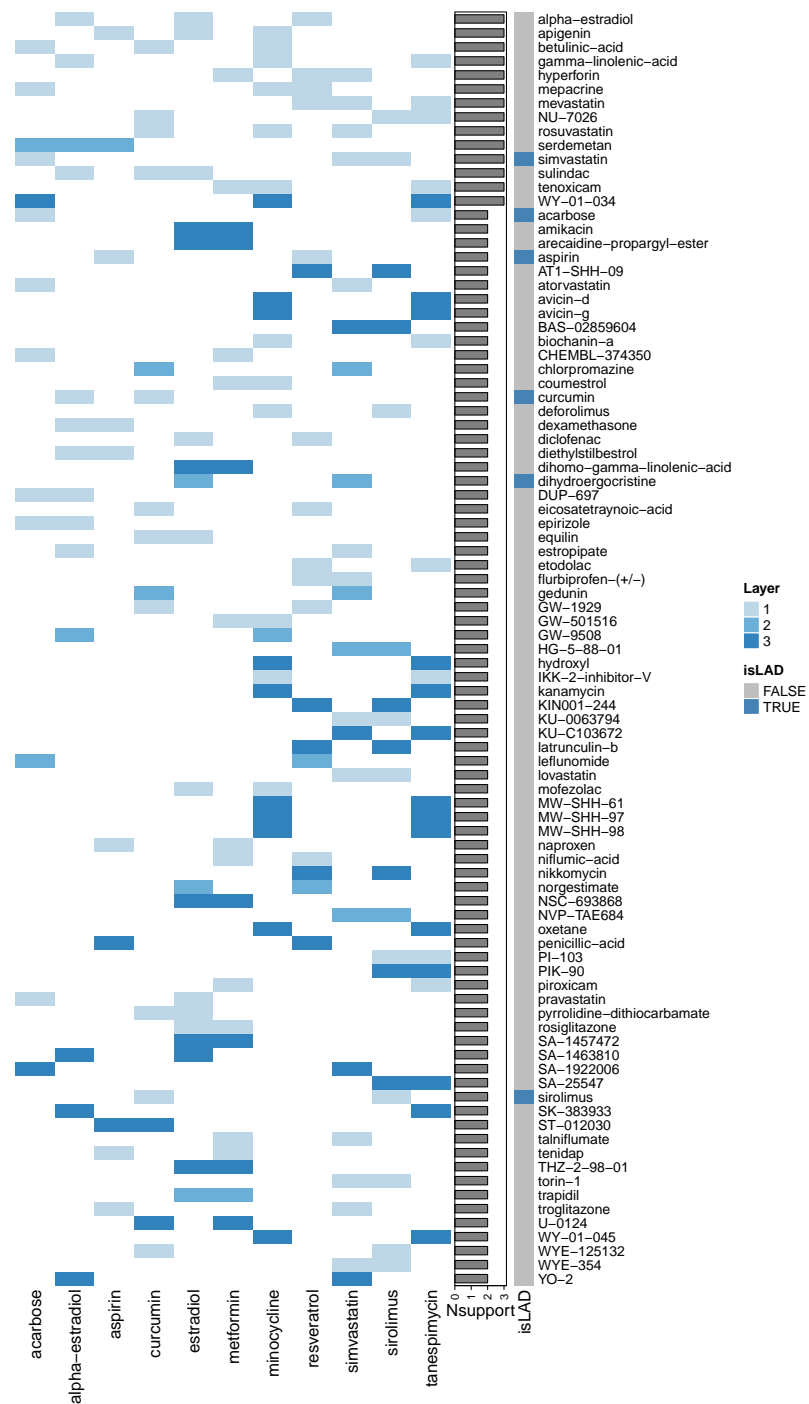


Figure 3.5: Combined PDs from LAD11 queries with at least 2 LADs support. The top 50 PDs from each individual LAD of LAD11 are combined into one drug list ranked by number of LADs queries that have the drug in their top 50 PD list (Nsupport) and the drugs are filtered with at least 2 Nsupport. It results in 91 drugs in this list. The corresponding table with no Nsupport cutoff is at Table S8.

Table 3.3: Top 10 ranking Reactome pathways in the FEA result from one sirolimus GES query in SKB cell with *dup_hyperG* method. N term/t/m: number of genes in the pathway, test set and intersect. P-adjust: P-value using the Benjamini-Hochberg (BH) method for multiple testing correction.

Annot	Term	N term/t/m	P-value	P-adjust
Reactome	Adrenoceptors (R-HSA-390696)	9/247/13	0.00e+00	0.00e+00
Reactome	Amine ligand-binding receptors (R-HSA-375280)	42/247/29	1.26e-38	5.79e-36
Reactome	Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	335/247/47	3.35e-24	1.03e-21
Reactome	GPCR ligand binding (R-HSA-500792)	467/247/47	4.37e-18	1.01e-15
Reactome	VEGFA-VEGFR2 Pathway (R-HSA-4420097)	99/247/21	5.89e-15	1.09e-12
Reactome	Signaling by VEGF (R-HSA-194138)	108/247/21	3.74e-14	5.74e-12
Reactome	CD28 co-stimulation (R-HSA-389356)	33/247/13	1.46e-13	1.92e-11
Reactome	CD28 dependent PI3K/Akt signaling (R-HSA-389357)	22/247/11	4.46e-13	5.05e-11
Reactome	G alpha (i) signalling events (R-HSA-418594)	405/247/37	4.94e-13	5.05e-11
Reactome	Nucleotide-like (purinergic) receptors (R-HSA-418038)	16/247/9	1.59e-11	1.47e-09

see whether the top ranking pathways in FEA results are correct target pathways of query drugs. For each LAD GES query, FEA analysis was performed on the targets of top 100 ranking drugs in the GESS results and the resulting pathways were ordered by adjusted p-values. Next, positions of target pathways of query drugs in their FEA results were plotted as dots in Figure 3.6A. Even though some drugs target a broad range of pathways, many of them matched in the top ranking positions in FEA results. In other words, the top ranking pathways in FEA results are those correctly targeted by query LADs, especially for LADs of sirolimus, metformin, tanespimycin, resveratrol, aspirin, alpha-estradiol and beta-estradiol.

Next, Reactome pathways in the FEA results were prioritized for each LAD across its multiple query GES samples in different cell types by voting strategy (Figure 3.6B, 3.6C and Figure S9-11, Table S11), the top 50 prioritized Reactome pathways (PPs) for each LAD were combined into one list by ranking them by the number of their support query LADs (Figure 3.7). Collectively, these results identified many Reactome pathways that might be related to human longevity with additional support on the enrichment results from genes in the [GeneAge](#) database (Table S10). The candidate LAPs include PI3K/AKT

Table 3.4: Summary table of LAD11 query GEs from in vitro and in vivo across different cell types or tissues. The starred cell types are used when the LAD names are used as sample names in the results. Condition: mouse condition when sacrificed, the numbers are age in month, M/F represent female/male, chronic means longer term treatment on old mouse.

SampleName	LAD	Source	Tissue	Cell	GEO Accession	Platform	Organism	Condition
acarbose_inVitro_muscle.SKB	acarbose	inVitro	muscle	SKB*		L1000	Human	
acarbose_inVitro_liver.PHH	acarbose	inVitro	liver	PHH		L1000	Human	
alpha-estradiol_inVitro_kidney.HA1E	alpha-estradiol	inVitro	kidney	HA1E*		L1000	Human	
aspirin_inVitro_breast.MCF7	aspirin	inVitro	breast	MCF7*		L1000	Human	
curcumin_inVitro_skin.FIBRNPC	curcumin	inVitro	skin	FIBRNPC*		L1000	Human	
estradiol_inVitro_kidney.HA1E	estradiol	inVitro	kidney	HA1E*		L1000	Human	
estradiol_inVitro_liver.HEPG2	estradiol	inVitro	liver	HEPG2		L1000	Human	
estradiol_inVitro_liver.PHH	estradiol	inVitro	liver	PHH		L1000	Human	
metformin_inVitro_skin.FIBRNPC	metformin	inVitro	skin	FIBRNPC*		L1000	Human	
metformin_inVitro_kidney.HEK293T	metformin	inVitro	kidney	HEK293T		L1000	Human	
minocycline_inVitro_breast.MCF7	minocycline	inVitro	breast	MCF7*		L1000	Human	
resveratrol_inVitro_muscle.SKB	resveratrol	inVitro	muscle	SKB*		L1000	Human	
resveratrol_inVitro_adipose.ASC	resveratrol	inVitro	adipose	ASC		L1000	Human	
resveratrol_inVitro_liver.PHH	resveratrol	inVitro	liver	PHH		L1000	Human	
simvastatin_inVitro_muscle.SKB	simvastatin	inVitro	muscle	SKB*		L1000	Human	
sirolimus_inVitro_muscle.SKB	sirolimus	inVitro	muscle	SKB*		L1000	Human	
sirolimus_inVitro_adipose.ASC	sirolimus	inVitro	adipose	ASC		L1000	Human	
sirolimus_inVitro_kidney.HA1E	sirolimus	inVitro	kidney	HA1E		L1000	Human	
sirolimus_inVitro_liver.HEPG2	sirolimus	inVitro	liver	HEPG2		L1000	Human	
sirolimus_inVitro_liver.PHH	sirolimus	inVitro	liver	PHH		L1000	Human	
tanespimycin_inVitro_skin.FIBRNPC	tanespimycin	inVitro	skin	FIBRNPC*		L1000	Human	
Acarbose12F_inVivo_Liver	acarbose	inVivo	Liver		GSE131754	RNAseq	Mouse	12F
Acarbose12M_inVivo_Liver	acarbose	inVivo	Liver		GSE131754	RNAseq	Mouse	12M
Acarbose6F_inVivo_Liver	acarbose	inVivo	Liver		GSE131754	RNAseq	Mouse	6F
Acarbose6M_inVivo_Liver	acarbose	inVivo	Liver		GSE131754	RNAseq	Mouse	6M
Estradiol6F_inVivo_Liver	estradiol	inVivo	Liver		GSE131754	RNAseq	Mouse	6F
Estradiol6M_inVivo_Liver	estradiol	inVivo	Liver		GSE131754	RNAseq	Mouse	6M
Metformin_inVivo_Adipose	metformin	inVivo	Adipose		GSE90755	RNAseq	Mouse	
Metformin_inVivo_Aorta	metformin	inVivo	Aorta		GSE90755	RNAseq	Mouse	
Metformin_inVivo_Brain	metformin	inVivo	Brain		GSE90755	RNAseq	Mouse	
Metformin_inVivo_Eyeball	metformin	inVivo	Eyeball		GSE90755	RNAseq	Mouse	
Metformin_inVivo_Heart	metformin	inVivo	Heart		GSE90755	RNAseq	Mouse	
Metformin_inVivo_Kidney	metformin	inVivo	Kidney		GSE90755	RNAseq	Mouse	
Metformin_inVivo_Liver	metformin	inVivo	Liver		GSE90755	RNAseq	Mouse	
Metformin_inVivo_Muscle	metformin	inVivo	Muscle		GSE90755	RNAseq	Mouse	
Metformin_inVivo_Stomach	metformin	inVivo	Stomach		GSE90755	RNAseq	Mouse	
Metformin_inVivo_Testis	metformin	inVivo	Testis		GSE90755	RNAseq	Mouse	
Resveratrol30M_inVivo_Heart	resveratrol	inVivo	Heart		GSE11291	Array	Mouse	30M
Resveratrol30M_inVivo_Muscle	resveratrol	inVivo	Muscle		GSE11291	Array	Mouse	30M
ResveratrolM_inVivo_Adipose	resveratrol	inVivo	Adipose		GSE11845	Array	Mouse	M
ResveratrolM_inVivo_Heart2	resveratrol	inVivo	Heart2		GSE11845	Array	Mouse	M
ResveratrolM_inVivo_Liver	resveratrol	inVivo	Liver		GSE11845	Array	Mouse	M
ResveratrolM_inVivo_Muscle2	resveratrol	inVivo	Muscle2		GSE11845	Array	Mouse	M
Sirolimus_inVivo_Adipose	sirolimus	inVivo	Adipose		GSE52825	Array	Mouse	
Sirolimus12F_inVivo_Liver	sirolimus	inVivo	Liver		GSE131754	RNAseq	Mouse	12F
Sirolimus12M_inVivo_Liver	sirolimus	inVivo	Liver		GSE131754	RNAseq	Mouse	12M
Sirolimus6F_inVivo_Liver	sirolimus	inVivo	Liver		GSE131754	RNAseq	Mouse	6F
Sirolimus6M_inVivo_Liver	sirolimus	inVivo	Liver		GSE131754	RNAseq	Mouse	6M
Sirolimus8M_inVivo_Liver	sirolimus	inVivo	Liver		GSE40977	Array	Mouse	8M
SirolimusChronicF_inVivo_Liver	sirolimus	inVivo	Liver		GSE48331	Array	Mouse	ChronicF
SirolimusChronicM_inVivo_Liver	sirolimus	inVivo	Liver		GSE48331	Array	Mouse	ChronicM
SirolimusF_inVivo_Heart	sirolimus	inVivo	Heart		GSE48043	RNAseq	Mouse	F
SirolimusF_inVivo_Liver	sirolimus	inVivo	Liver		GSE48331	Array	Mouse	F
SirolimusM_inVivo_Heart	sirolimus	inVivo	Heart		GSE41018	Array	Mouse	M
SirolimusM_inVivo_Kidney	sirolimus	inVivo	Kidney		GSE41018	Array	Mouse	M
SirolimusM_inVivo_Liver	sirolimus	inVivo	Liver		GSE48331	Array	Mouse	M

Table 3.5: The manually curated annotations for all of the prioritized DrugAge drugs (Pri-oDA) in this project including drug description, tested assays, publications, *etc.* PCID: PubChem CID. The elaborate information is at Table S16.

PrioDA	Description	species	strain	dosage	avg lifespan change	gender	pubmed id	PCID
wortmannin	fungal metabolite identified as a potent and selective inhibitor for PI3Ks	Caenorhabditis elegans; Drosophila melanogaster	N2; Canton S	1 μ M; 5 μ M	-4.1; 5	Male	23543623; 24096697	312145
cinnarizine	antihistamine used for motion sickness	Caenorhabditis elegans		33 μ M	15		24134630	1547484
SU-4312	selective and potent VEGFR inhibitor	Caenorhabditis elegans		88 μ M	5		24134630	6450842
artigenin	plant lignan with antioxidant, anti-inflammatory, anti-cancer and antiviral activities	Caenorhabditis elegans	N2; N2	10 μ M; 100 μ M	6.7; 13.7		26141518; 26141518	64981
dihydroergocristine	ergot alkaloid with partial agonist activity on dopaminergic and alpha-adrenergic receptors and antagonist activity on serotonin receptors. Approved by FDA for treatment of dementia like Alzheimer's	Caenorhabditis elegans		176 μ M	34		24134630	444034
lonidamine	interferes with energy metabolism of cancer cells, inhibit aerobic glycolytic activity on mitochondrially-bound hexokinase (HK)	Caenorhabditis elegans		5 μ M	8		21932172	39562
quercetin	plant pigment that belongs to flavonoids	Mus musculus; Caenorhabditis elegans; Aedes aegypti; Drosophila melanogaster; Podospora anserina; Aedes albopictus	LACA; N2; Liverpool; Canton S	10 mg/day; 100 μ M; 200 μ M; 22 μ g/mL; 250 μ M; 50 μ M; 0.3 μ M; 0.5 μ M; 1 μ M; 300 μ M; 50 ppm; 100 ppm; 200 ppm	-5.8; -14.2; -11.6; 19; 10; 10; 8; 8; 9; 10; 10; 9; 10; 11; 18; 11; 14; 10.82; 5.79; -7; 2; 9; 11; 10; 59.85; 30; -2.8; -12; -13.79; -5.09; -4.55; -10.2; 13; 13; 10.2; 19.01; 20.88; 20.53; 15; -3; 16; -8; 0; -2; 25; 29; 9; 5.11; 18.35; 4.76; 3.4; -13.07; -13.84; -14.62; -8.2; -6.7; -8.21; 6; 16.6	FEMALE; MALE; BOTH	7140862; 18024103; 18692520; 19043800; 21563825; 21776484; 22155175; 22493606; 27732590; 28066251; 29780405; 34188892	5280343
geldanamycin	antitumor antibiotic that potently inhibits the function of HSP90 that play important roles in regulation of cell cycle, cell growth, apoptosis, angiogenesis and oncogenesis.	Caenorhabditis elegans; Drosophila melanogaster	N2; Canton-S	20 μ M; 10 μ M; 100 μ M; 200 μ M	2; 3.51; 0.84; 0.22; 5.56	Female; Male	26676933; 33008901	5288382
CGP-52411	selective inhibitor of EGFR, also inhibits and reverses formation of Abeta42 fibers associated with Alzheimer's	Caenorhabditis elegans		13 μ M	15		24134630	1697
staurosporine	alkaloid isolated from bacteria exhibiting anti-cancer activity. Potent inhibitor of protein kinases	Drosophila melanogaster	Oregon R	30-50 μ M	34.8		22363408	44259
caffeine	methylxanthine alkaloid, acts primarily as an adenosine receptor antagonist in the central nervous system (CNS) with psychotropic and anti-inflammatory activities	Drosophila melanogaster; Saccharomyces cerevisiae; Caenorhabditis elegans; Drosophila melanogaster; Aedes albopictus	Oregon R; BY4741; N2; Canton-S	0.01 mg/ml; 1 mg/ml; 0.1 mg/ml; 0.2-0.4 mM; 100 mM; 75 mM; 7.5 mM; 30 mM; 5 mM; 0.5 mM; 10 mM; 2.5 mM; 20 mM; 50 mM; 60 mM; 45 mM; 0.03%; 0.05%; 50 μ g/mL; 1 mM; 50 ppm; 100 ppm; 200 ppm	-10.1; -3.3; 1.7; 86; -89.5; -82.8; -11.6; -24.4; -13.6; 7.7; 8.7; 16.9; 17.2; 19.6; 21.6; -40.9; 10.8; 36.7; 28.5; -96; -92; -60; 0; 24; 8; 35; 65; -22.86; -17.14; 15.39; 31.9; 16.42; 0.1; -4.64; -6.42; -11.6; -11.6	Male; FEMALE; MALE	8326745; 18513215; 24764514; 26696878; 29093334; 30061824; 31432005; 34188892	2519
pyrazolanthrone	anthrone derivative, inhibitor of c-Jun N-terminal kinases (JNKs) (Bennett et al. 2001)	Drosophila melanogaster	Oregon R	10 mM	23.5		22363408	8515

related pathways (PI3K/AKT signaling in cancer, PIP3 activates AKT signaling, activated NTRK3 signals through PI3K, PI5P, PP2A, and IER3 regulate PI3K/AKT Signaling), signaling by VEGF, regulation of PTEN gene transcription, regulation of TP53 degradation, insulin receptor signaling cascade, immune system related pathways (signaling by interleukins, interleukin-4 and interleukin-13 signaling), SUMOylation of intracellular receptors, extra-nuclear estrogen signaling from sirolimus query. VEGFA-VEGFR2 (vascular endothelial cell growth factor receptor) pathway, signaling by ERBB4, cytochrome P450 - arranged by substrate type, post NMDA receptor activation events, extra-nuclear estrogen signaling, MAP kinase activation, xenobiotics from acarbose query. Pathways like xenobiotics (commonly affected pathway for organisms treated with drugs), GPCR ligand binding, neuron system related (neurotransmitter receptors and postsynaptic signal transmission, activation of NMDA receptors and postsynaptic events, dopamine receptors), SUMOylation of intracellular receptors, extra-nuclear estrogen signaling, intracellular signaling by second messengers, PI3K/AKT signaling in cancer, VEGFA-VEGFR2 pathway, RAS signaling (CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signaling, signaling by RAS mutants), G alpha (i) signalling events, MAPK family signaling cascades, signaling by RAF1 mutants are supported by multiple LADs. Reactome enrichment analysis was also performed on the top 50 PDs from sirolimus query (Table S12) and compared with its top 50 PPs in Figure 3.6B. The top 50 ranking terms from the two Reactome pathway sets have a very good consistency (32/50 overlapped pathways, Table 3.6), making the results more robust and trustworthy. The overlapped Reactome pathways include PI3K/AKT related pathways (MET activates PI3K/AKT signaling, PIP3 activates AKT signaling), intracellu-

lar signaling by second messengers, signaling by VEGF, diseases of signal transduction by growth factor receptors and second messengers, signaling by PDGFRA extracellular domain mutants, regulation of TP53 degradation, insulin receptor signalling cascade, RET signaling, signaling by erythropoietin, and signaling by type 1 Insulin-like Growth Factor Receptor (IGF1R).

In summary, the PPs are mainly involved in PI3K/AKT related pathways, estrogen and steroid related pathways, neurotransmitter signaling, immune system, MAPK signaling, RAS signaling, RAF1 signaling, AMPK signaling [20, 207, 222], vascular endothelial growth factor (VEGF) related signaling, insulin signaling, TP53 regulations, and G alpha (i) signalling. Many studies have demonstrated that the downregulation of signaling pathway of insulin/insulin-like growth factor-1 (IGF-1)/phosphatidylinositol-3 kinase (PI3K)/Akt can extend longevity as well as resistance to oxidative stress in the nematode *Caenorhabditis elegans* [70, 139, 131, 113, 118, 98]. Akt negatively regulates the in vitro lifespan of human endothelial cells via a p53/p21-dependent pathway, inhibition of Akt extends the lifespan of endothelial cells [129]. G alpha (i) signalling is a ubiquitously expressed pathway that makes up a broad class of signal transduction cascades involving interactions with G protein coupled receptors (GPCR). They are assembled heterogeneously to regulate both cellular and physiological functions. On the cellular level, G alpha (I) signaling regulates a variety of functions including the regulation of K^+ channels [141], as well as cell proliferation via ERK and RAS signaling [64]. On the organismal level, G alpha (i) signalling has a neuron, neuroendocrine, and platelet preferential expression pattern. Within this expression profile, G alpha (I) functions downstream of beta2 adrenergic receptors thereby promoting

parasympathetic tone, which may be important in the prevention of aging-related elevated blood pressure [5, 210]. These pathways are promising for human healthy aging and may be pharmacologically targeted to extend lifespan.

The signaling by VEGF (R-HSA-194138) pathway was then selected based on its overall high rankings as well as its potential as a LAP and a drug-target (DT) interaction network was created to resolve the relationships between the drugs in GESS results and their molecular targets (Figure 3.6D, Table 3.7). Within this network, several well-known LADs (sirolimus, wortmannin, caffeine, tanespimycin, everolimus) and LAGs (MTOR, PIK3CA, PIK3R1, AKT, MAPK14, HSP90AA1, VEGFA) were identified. It also identifies new potential LADs including QL-X-138 (BTK/MNK Dual Inhibitor for Lymphoma and Leukemia [218]), CGP-53353, perphenazine, crizotinib, purvalanol-a, GSK-1059615, lovastatin, KU-0063794 (an MTOR inhibitor) and tamoxifen, new potential LAGs including PRKCA, CALM1, PAK3, PTK2, ROCK1, ITPR1, FLT1, RHOA, CDH5, PAK2 and PRKCZ. KU-0063794 is also identified by Tyshkovskiy *et al.* (2019) [195] as new lifespan-extending candidates, it adds convincing to the results.

3.2.6 GESS/FEA Results for Longevity-based Phenotype Signature

In addition to drug induced GESs, the GESS/FEA workflow was also applied to a human phenotype-based GES to discover candidate LADs and LAPs (Figure 3.1F). This was done to both validate the results obtained from drug-based signatures above and identify new ones that are unique to this query type. Peters *et al.* (2015) [145] identified 1,497 DEGs associated with chronological age from whole-blood gene expression meta-analysis. The up- and down-regulated 150 genes were chosen as the query GES for the GESS/FEA workflow

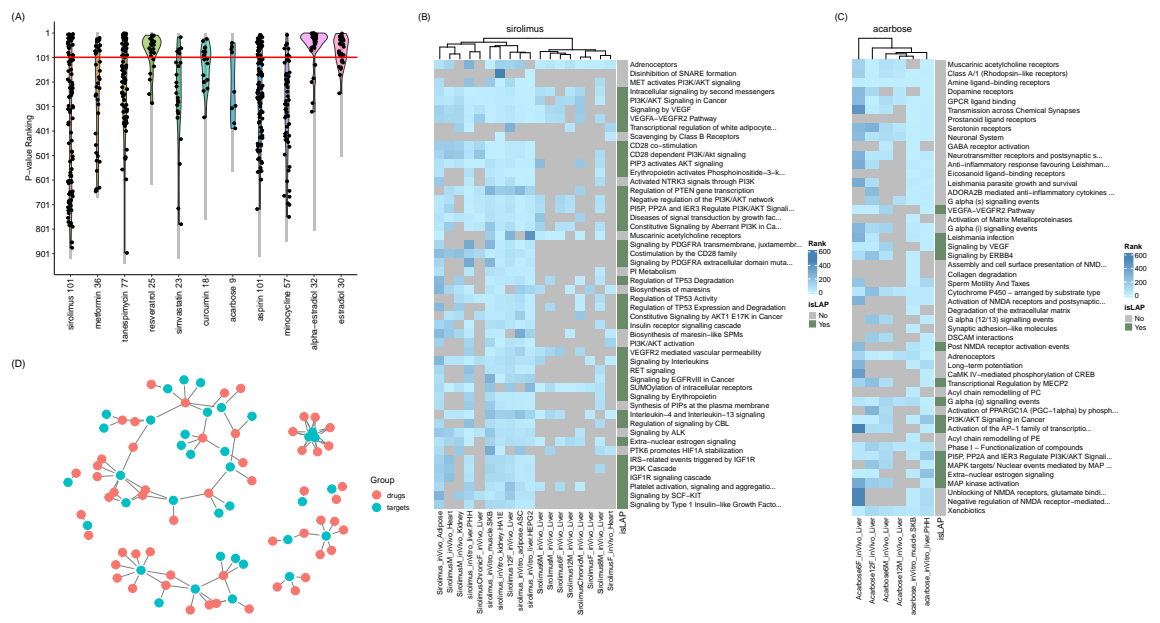


Figure 3.6: FEA results summary. A: ranking positions of Reactome pathways meeting a 0.05 adjusted p-value cutoff from targets of the query LADs in their FEA results. Black dots represent pathways in FEA results matching the direct pathways from the query LAD targets. Grey bars indicate the total number of pathways in FEA results. B and C: Top 50 PPs from vote strategy on query LADs of sirolimus and acarbose, respectively. The columns are query samples that are clustered by Euclidian distance of color key (rankings transformed from adjusted p-values). Known LAPs are annotated in green in the binary color bar to the right of the heatmap. The complete tables containing all Reactome pathway rankings from voting strategy with scores and LAP annotation corresponding to the heatmaps are stored in Synapse ([syn27074585](https://synapse.org/syn27074585)). D: DT network of the signaling by VEGF (R-HSA-194138) reactome pathway. Symbols of drugs, targets and their relationship are available at Table 3.7.

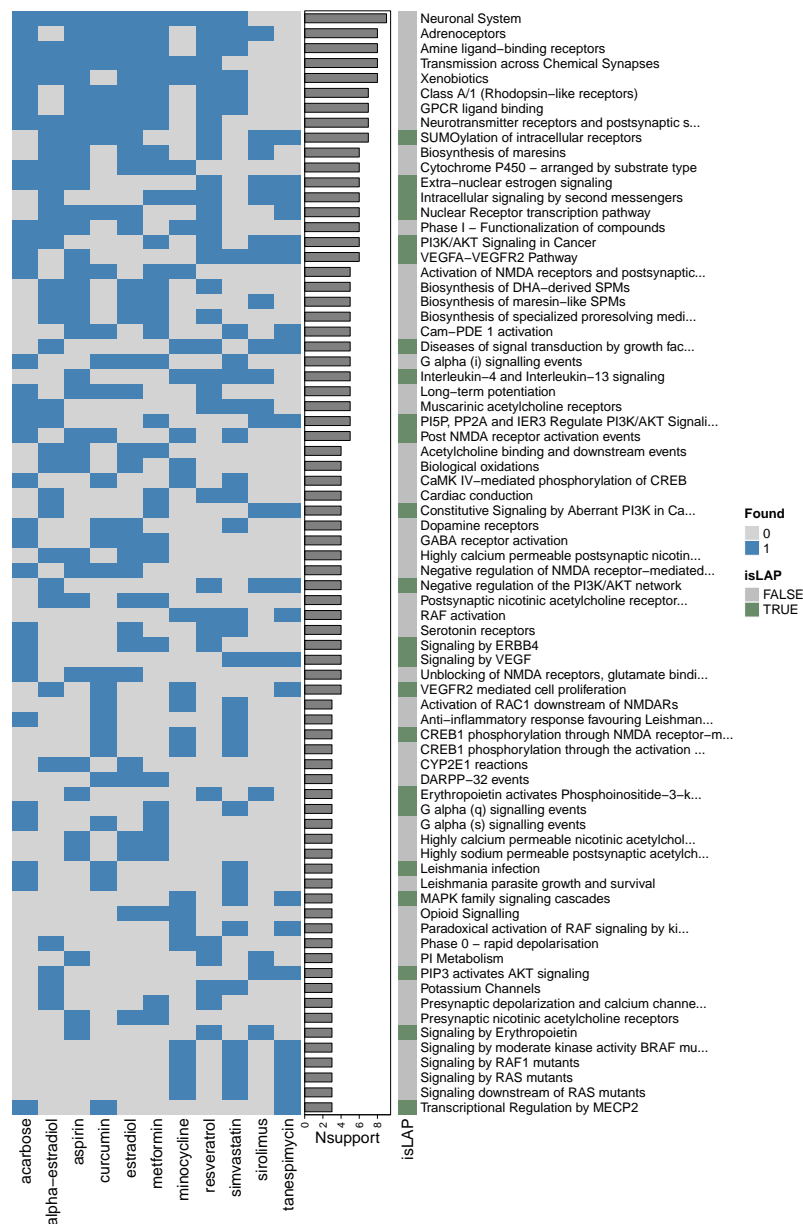


Figure 3.7: Combined PPs from LAD11 queries with at least 3 LADs support. The top 50 PPs from each LAD of LAD11 are combined into one final list ranked by the number of query LADs that have the pathway in its top 50 PPs (Nsupport) and filtered by selecting those that have at least 3 Nsupport. It results in 74 Reactome pathways in this list. The corresponding table with no Nsupport cutoff is at Table S11.

Table 3.6: Overlapped Reactome pathways between top 50 enriched terms from Sirolimus top 50 prioritized drug list and top 50 prioritized list in Sirolimus FEA results.

ID	Description	isLAP
R-HSA-8851907	MET activates PI3K/AKT signaling	FALSE
R-HSA-9006925	Intracellular signaling by second messengers	TRUE
R-HSA-2219528	PI3K/AKT Signaling in Cancer	TRUE
R-HSA-194138	Signaling by VEGF	TRUE
R-HSA-4420097	VEGFA-VEGFR2 Pathway	TRUE
R-HSA-389356	CD28 co-stimulation	TRUE
R-HSA-389357	CD28 dependent PI3K/Akt signaling	TRUE
R-HSA-1257604	PIP3 activates AKT signaling	TRUE
R-HSA-9027276	Erythropoietin activates Phosphoinositide-3-kinase (PI3K)	TRUE
R-HSA-9603381	Activated NTRK3 signals through PI3K	FALSE
R-HSA-199418	Negative regulation of the PI3K/AKT network	TRUE
R-HSA-6811558	PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling	TRUE
R-HSA-5663202	Diseases of signal transduction by growth factor receptors and second messengers	TRUE
R-HSA-2219530	Constitutive Signaling by Aberrant PI3K in Cancer	TRUE
R-HSA-9673767	Signaling by PDGFRA transmembrane, juxtamembrane and kinase domain mutants	TRUE
R-HSA-388841	Costimulation by the CD28 family	TRUE
R-HSA-9673770	Signaling by PDGFRA extracellular domain mutants	TRUE
R-HSA-1483255	PI Metabolism	FALSE
R-HSA-6804757	Regulation of TP53 Degradation	TRUE
R-HSA-5633007	Regulation of TP53 Activity	TRUE
R-HSA-6806003	Regulation of TP53 Expression and Degradation	TRUE
R-HSA-5674400	Constitutive Signaling by AKT1 E17K in Cancer	TRUE
R-HSA-74751	Insulin receptor signalling cascade	TRUE
R-HSA-198203	PI3K/AKT activation	FALSE
R-HSA-5218920	VEGFR2 mediated vascular permeability	TRUE
R-HSA-8853659	RET signaling	TRUE
R-HSA-9006335	Signaling by Erythropoietin	TRUE
R-HSA-1660499	Synthesis of PIPs at the plasma membrane	FALSE
R-HSA-912631	Regulation of signaling by CBL	TRUE
R-HSA-2428928	IRS-related events triggered by IGF1R	TRUE
R-HSA-2428924	IGF1R signaling cascade	TRUE
R-HSA-2404192	Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)	TRUE

Table 3.7: Drugs and targets involved in the DT network of signaling by VEGF (R-HSA-194138) in Figure 3.6D.

Drug	Targets
sirolimus	MTOR
wortmannin	MTOR; PIK3CA; PIK3R1
phorbol-12-myristate-13-acetate	PRKCA
QL-X-138	MTOR
AZD-8055	MTOR
perphenazine	CALM1; CALM2; CALM3
WYE-354	MTOR
BMS-754807	AKT1
torin-1	MTOR; PIK3CA
H-7	PRKACA
BX-795	KDR
NVP-TAE684	AXL; PAK3; PTK2; PTK2B; ROCK1
caffeine	ITPR1; ITPR2; ITPR3; PIK3CA; PIK3CB
CGP-53353	PRKCB
JX-401	MAPK14
tanespimycin	HSP90AA1
geldanamycin	HSP90AA1
NVP-AUY922	HSP90AA1
AS-605240	PIK3CA; PIK3CB
crizotinib	AXL
RHO-kinase-inhibitor-III[rockout]	ROCK1
purvalanol-a	SRC
tricitriline	AKT1; AKT2; AKT3
GSK-1059615	PIK3CA
PU-H71	HSP90AA1
U-0126	AKT1; MAPK11; MAPK12; MAPK14; PRKCA; ROCK1
BMS-536924	AKT1; KDR
felodipine	CALM1; CALM2; CALM3
KI-8751	KDR
cediranib	FLT1; FLT4; KDR
atorvastatin	RHOA
chlorpromazine	CALM1; CALM2; CALM3
gedunin	HSP90AA1
lovastatin	RHOA
KU-0063794	MTOR
lenalidomide	CDH5
cytochalasin-b	ACTB
H-89	PRKACA
quizartinib	FLT4
KU-0060648	PIK3CA; PIK3CB
SU-4312	KDR
TGX-221	PIK3CB
pimozide	CALM1; CALM2; CALM3
motesanib	FLT1; FLT4; KDR
nicardipine	CALM1; CALM2; CALM3
quercetin	ACTB; HSP90AA1
carvedilol	VEGFA
bucladesine	PRKACA
everolimus	MTOR
minocycline	VEGFA
SU-11652	FLT1; KDR
staurosporine	FLT4; MAPKAPK2; PAK2; PRKCB
cinchocaine	CALM1; CALM2; CALM3
semaxanib	FLT1; KDR
nifedipine	CALM1; CALM2; CALM3
tamoxifen	PRKCA; PRKCB; PRKCZ

to identify drugs that induce GESs that are positively and/or negatively correlated with this longevity-associated signature. The same GESS and FEA methods were used here as the LAD11 queries. The NCS score distributions from Peters *et al.* (2015) query are more significant when compared to a randomly generated signature (Figure 3.8A) but not when compared to the GESS results from drug treatment (Figure 3.4A). Following GESS, RRSs are calculated at 5, 10 and 15 percent randomization, ranked from largest to lowest by NCS scores, and the top 10 positively connected drugs are shown in Table 3.8. Although RRSs for this query are more moderate compared to the GESS results obtained from the drugs (Table 3.2), The most highly positively ranking and interesting drugs (with $RRS > 0.5$) include BRD-K63954456 (PubChem CID 2202512, a macrophage migration inhibitory factor) and erlotinib (PubChem CID 176871, kinase inhibitor that helps slow or stop the spread of cancer cells by blocking the action of an abnormal protein that signals cancer cells to multiply). Out of the top 500 ranking drugs and small molecules in Peters GESS results, 52 drugs are overlapped with the combined PDs from LAD11 (Table 3.9). It give more credibility to the overlapped drugs that can be assume to be related to human longevity and healthy aging including pravastatin, amsacrine, thioridazine, exemestane, rosuvastatin, chlorphenamine, gemfibrozil, crizotinib, alprostadil, meloxicam, mepacrine, GW-501516 (also known as cardarine, a PPAR receptor agonist that was developed as a drug candidate for metabolic and cardiovascular diseases, but was abandoned in 2007 because animal testing showed that the drug caused cancer to develop rapidly in several organs) that have therapeutic efficacy or under FDA study. The known LADs or DrugAge drugs of sirolimus, simvastatin, KN-93, phenformin, pyrazolanthrone, arctigenin, caffeine, wortmannin are also ranked among the

top 500s in the Peters GESS results indicating that they induce similar GESSs to Peters *et al.* (2015) query (Figure 3.8B).

Next, FEA analysis was performed on the top 100 drugs in the Peters GESS result using the *dup_hyperG* method followed by computing RRSs to get the top ranking KEGG and Reactome pathways (Table 3.10, Table S15). Three out of ten KEGG pathways have RRS5 greater than 0.3, including EGFR tyrosine kinase inhibitor resistance (EGFR-TKIR), ErbB signaling pathway and Glioma. The epidermal growth factor receptor (EGFR) is a receptor that, when activated, promotes cell proliferation [204]. Genetic studies have identified several activating mutations within the EGFR coding sequence that are believed to be causative to non small cell lung cancer (NSCLC) [121]. Although EGFR tyrosine kinase inhibitors are effective at treating NSCLC, a number of tumors develop EGFR-TKIR and are difficult to treat [105]. Yet, pharmacological studies suggest that treating EGFR-TKIR with longevity promoting compounds including vorinostat, metformin, or resveratrol are effective at reversing EGFR-TKIR suggesting that pharmacologically inhibiting EGFR or its downstream signaling events may have longevity promoting effects [29, 230]. ErbB receptors are a class of receptor that also activate a number of cellular proliferation pathways including those that, when inhibited, promote longevity such as the MAPK and AKT-PI3K-mTOR pathways [207, 222, 212]. The Reactome pathway results also include several pathways related to longevity in the top 20 hits including diseases of signal transduction by growth factor receptors and second messengers, PI3K/AKT signaling in cancer, neurotransmitter receptors and postsynaptic signal transmission (Table S15). Out of the top 100 ranking Reactome pathways in Peters FEA results, about half of them (52) are overlapped

with the combined PPs from LAD11 (Table 3.11). It adds more evidence to the identified pathways that might be related to human longevity including growth factor receptor related pathways (diseases of signal transduction by growth factor receptors and second messengers, VEGFA-VEGFR2 pathway, signaling by VEGF, VEGFR2 mediated vascular permeability), PI3K/AKT related signaling pathways, neuron system (neurotransmitter receptors and postsynaptic signal transmission, neuronal system), metabolism of steroids and estrogen signaling, immune system (interleukin-4 and interleukin-13 signaling), regulation of PTEN gene transcription, MAPK, ERBB4 and G alpha (q) signalling.

Next, a DT network was created for the EGFR-TKIR pathway in order to further resolve the relationships between the selected drugs and their molecular targets (Figure 3.8C). Several well-studied drugs were identified within this target network including sirolimus, a well-known LAD, as well as 10 additional drugs (erlotinib, PLX-4720, hispidin, AG-957, mepacrine, vemurafenib, atorvastatin, GSK-3-inhibitor-IX, oxindole-I, gefitinib) with comparable NCS scores to sirolimus that may be candidate LADs (Table 3.12). In addition to putative LADs, a number of genes within the EGFR-TKIR pathway are inhibited by drugs that promote longevity including EGFR, MTOR, RAF1 and MAPK3 [159, 175, 207]. Thus, it is likely that the development of inhibitors that target additional EGFR-TKIR pathway members may have lifespan extending effects. In summary, these results identified a number of drugs, genes and pathways that promote longevity or prevent aging related diseases.

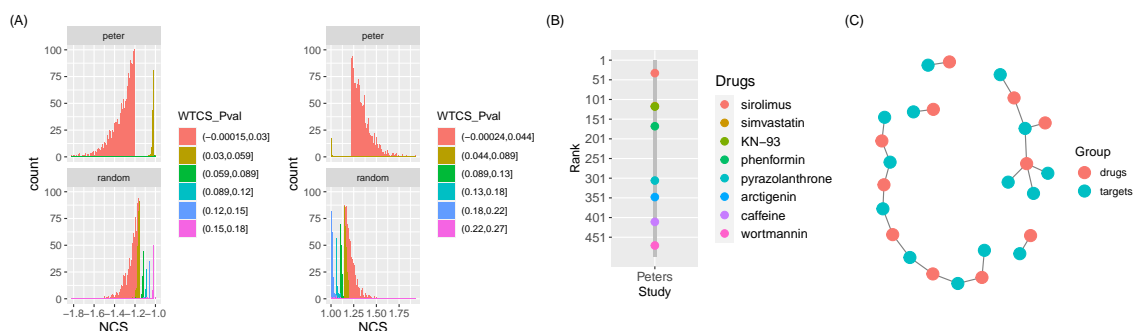


Figure 3.8: Peters *et al.* (2015) query results. A: NCS score distributions in the GESS results from Peters *et al.* (2015) query and one random GES query as negative control after setting count cutoff as 100, the left and right panels show the negative part and positive part of NCS scores, respectively. The color key shows the P-values of the WTCS score for the entries in the GESS results. B: Position of known LADs from DrugAge database in Peters GESS result. Due to space limitation, it only shows the position of LADs in the top 500 drugs whose GESs are positively connected with the query GES out of 8140. C: DT network for the EGFR tyrosine kinase inhibitor resistance (EGFR-TKIR) pathway. Symbols of drugs, targets and their relationship are available at Table 3.12.

Table 3.8: Top 10 ranking positively connected drugs of Peters GESS result from *LINCS* method. Cell: cell type, NEU: normal cell, PC3: prostate adenocarcinoma, VCAP: prostate carcinoma, ASC: normal primary adipocyte stem cells; Targets: gene symbol of protein targets; RRS5,10,15: rank robustness scores at 5, 10, 15 percent randomization. For detailed description of the score columns for LINCS method, please consult to the vignette of the *signatureSearch* package. The complete table is available at Table S13.

Drug Name	Cell	NCS	Tau	Targets	RRS5	RRS10	RRS15
BRD-K63954456	NEU	1.93	99.96		0.97	0.90	0.82
pravastatin	SKB	1.83	99.79	ABCC2; APOB; CCL2; CETP; CRP; HMGCR; LCAT; SELP; SLCO1B1	0.15	0.22	0.22
erlotinib	ASC	1.78	99.87	EGF; EGFR; GAK; MAP3K19; NR1I2; SLK; STK10	0.61	0.44	0.31
calcitriol	HL60	1.78		CDA; CYP24A1; CYP27B1; CYP3A5; GC; HOXA10; VDR	0.13	0.21	0.23
BRD-A07614565	SKB	1.74	99.83		0.13	0.10	0.10
AG-957	VCAP	1.72	98.47	ABL1; EGFR	0.02	0.04	0.03
BRD-K12411950	MCF7	1.72	99.92		0.19	0.12	0.09
SA-90377	NPC	1.72	99.64		0.05	0.07	0.10
BRD-K42021584	PC3	1.71	99.48		0.01	0.02	0.01
BRD-K37762845	ASC	1.71	99.91		0.18	0.11	0.09

Table 3.9: Overlapped drugs between top 500 positively connected drugs in Peters GESS results and combined PDs summarized from LAD11. Therapeutic: whether the drugs have therapeutic effect. Max Phase: FDA max phase study.

Drug	isLAD	Therapeutic	Max Phase
pravastatin	FALSE	0	0
quinoclamine	FALSE	0	0
SA-1921085	FALSE		
NU-7026	FALSE	0	0
mepacrine	FALSE	0	2
sirolimus	TRUE		
atorvastatin	FALSE		
mevastatin	FALSE		
lovastatin	FALSE	0	0
GW-9508	FALSE	0	0
SB-203186	FALSE	0	0
brazilin	FALSE	0	0
amsacrine	FALSE	1	4
simvastatin	TRUE	1	4
thioridazine	FALSE	1	4
biochanin-a	FALSE	0	0
manumycin-a	FALSE		
exemestane	FALSE	1	4
CHEMBL-374350	FALSE	0	0
rosuvastatin	FALSE	1	4
ricinine	FALSE	0	0
DY-44	FALSE		
HG-5-88-01	FALSE	0	0
PD-407824	FALSE	0	0
QL-X-138	FALSE	0	0
SA-1921456	FALSE		
CGS-12066B	FALSE		
alprenolol	FALSE	0	0
fluperlapine	FALSE	0	0
chlorphenamine	FALSE	1	4
gemfibrozil	FALSE	1	4
crizotinib	FALSE	1	4
calcifediol	FALSE		
torin-1	FALSE	0	0
PP-3	FALSE	0	0
WZ-7043	FALSE		
tubastatin-a	FALSE	0	0
KUC111774N	FALSE	0	0
arctigenin	TRUE	0	0
norethisterone	FALSE	0	0
GW-501516	FALSE	0	2
AS-605240	FALSE	0	0
STOCK1S-03920	FALSE	0	0
alprostadil	FALSE	1	4
WH-4023	FALSE	0	0
meloxicam	FALSE	1	4
chlorpromazine	FALSE		
LY-278584	FALSE	0	0
arvanil	FALSE	0	0
wortmannin	TRUE	0	0
prostaglandin	FALSE		
tremulacin	FALSE	0	0

Table 3.10: Top 10 ranking KEGG pathways from Peters *et al.* (2015) longevity-based query GES. RRS5, 10, 15: rank robustness scores at 5, 10 and 15 percent randomization. The complete table is available at Table S14.

ID	Description	GeneRatio	BgRatio	pvalue	RRS5	RRS10	RRS15
hsa01521	EGFR tyrosine kinase inhibitor resistance	20/168	79/7444	3.2e-16	0.56	0.46	0.40
hsa05212	Pancreatic cancer	17/168	75/7444	4.3e-13	0.01	0.02	0.02
hsa05205	Proteoglycans in cancer	25/168	203/7444	2.6e-12	0.04	0.03	0.02
hsa04012	ErbB signaling pathway	17/168	85/7444	3.8e-12	0.35	0.36	0.35
hsa05210	Colorectal cancer	17/168	86/7444	4.7e-12	0.11	0.10	0.08
hsa04726	Serotonergic synapse	19/168	115/7444	7.0e-12	0.20	0.18	0.17
hsa04979	Cholesterol metabolism	12/168	50/7444	6.8e-10	0.20	0.16	0.14
hsa05214	Glioma	14/168	75/7444	8.8e-10	0.47	0.32	0.30
hsa04014	Ras signaling pathway	23/168	232/7444	1.7e-09	0.21	0.21	0.18
hsa04270	Vascular smooth muscle contraction	17/168	127/7444	2.8e-09	0.13	0.12	0.10

3.3 Discussion

In this project, several longevity associated MOAs were found from GESs queries of LAD87 and LAD11 (Figure 3.2B and 3.2C). It prioritized several MOAs that are functionally related to healthy aging and longevity including chloride channel blocker, tricyclic antidepressant, MAP kinase inhibitor, PPAR receptor agonist, RAF inhibitor, TGF beta receptor inhibitor, glycogen synthase kinase inhibitor, anthelmintic, dopamine uptake inhibitor, retinoid receptor agonist, tachykinin antagonist, sigma receptor antagonist, casein kinase inhibitor, PI3K inhibitor, tubulin inhibitor. The identified MOAs have a large scope of affected biological systems and pathways, indicating that human longevity is a complex phenotype and involved in a lot of pathways that maybe pharmacological targeted to promote longevity and healthy aging. The target genes/proteins of LAD11 were mainly mapped to Reactome pathways of SUMOylation (part of metabolism of proteins); activation of AMPK downstream of NMDARs (part of neuronal system); activation of PPARGC1A and mitochondrial biogenesis (part of organelle biogenesis and maintenance); MTOR signaling and Extra-nuclear estrogen signaling (part of signal transduction); translocation of SLC2A4

Table 3.11: Overlapped Reactome pathways between top 100 terms in Peters FEA results and combined PPs summarized from LAD11.

ID	Description	isLAP
R-HSA-5663202	Diseases of signal transduction by growth factor receptors and second messengers	TRUE
R-HSA-375280	Amine ligand-binding receptors	FALSE
R-HSA-8963899	Plasma lipoprotein remodeling	TRUE
R-HSA-8848021	Signaling by PTK6	TRUE
R-HSA-9006927	Signaling by Non-Receptor Tyrosine Kinases	TRUE
R-HSA-2219528	PI3K/AKT Signaling in Cancer	TRUE
R-HSA-1257604	PIP3 activates AKT signaling	TRUE
R-HSA-174824	Plasma lipoprotein assembly, remodeling, and clearance	FALSE
R-HSA-9006925	Intracellular signaling by second messengers	TRUE
R-HSA-112314	Neurotransmitter receptors and postsynaptic signal transmission	FALSE
R-HSA-8964026	Chylomicron clearance	FALSE
R-HSA-6811558	PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling	TRUE
R-HSA-199418	Negative regulation of the PI3K/AKT network	TRUE
R-HSA-112315	Transmission across Chemical Synapses	FALSE
R-HSA-8963888	Chylomicron assembly	FALSE
R-HSA-8963901	Chylomicron remodeling	FALSE
R-HSA-1227986	Signaling by ERBB2	TRUE
R-HSA-211897	Cytochrome P450 - arranged by substrate type	FALSE
R-HSA-8957322	Metabolism of steroids	FALSE
R-HSA-211981	Xenobiotics	FALSE
R-HSA-390696	Adrenoceptors	FALSE
R-HSA-211945	Phase I - Functionalization of compounds	FALSE
R-HSA-76002	Platelet activation, signaling and aggregation	TRUE
R-HSA-2219530	Constitutive Signaling by Aberrant PI3K in Cancer	TRUE
R-HSA-8963898	Plasma lipoprotein assembly	TRUE
R-HSA-9009391	Extra-nuclear estrogen signaling	TRUE
R-HSA-8943724	Regulation of PTEN gene transcription	TRUE
R-HSA-4420097	VEGFA-VEGFR2 Pathway	TRUE
R-HSA-3000471	Scavenging by Class B Receptors	FALSE
R-HSA-5674499	Negative feedback regulation of MAPK pathway	FALSE
R-HSA-8857538	PTK6 promotes HIF1A stabilization	FALSE
R-HSA-194138	Signaling by VEGF	TRUE
R-HSA-6785807	Interleukin-4 and Interleukin-13 signaling	TRUE
R-HSA-9006931	Signaling by Nuclear Receptors	TRUE
R-HSA-5674400	Constitutive Signaling by AKT1 E17K in Cancer	TRUE
R-HSA-8866910	TFAP2 (AP-2) family regulates transcription of growth factors and their receptors	TRUE
R-HSA-112316	Neuronal System	FALSE
R-HSA-1236394	Signaling by ERBB4	TRUE
R-HSA-9018682	Biosynthesis of maresins	FALSE
R-HSA-975634	Retinoid metabolism and transport	FALSE
R-HSA-373076	Class A/1 (Rhodopsin-like receptors)	FALSE
R-HSA-9018678	Biosynthesis of specialized proresolving mediators (SPMs)	FALSE
R-HSA-6806667	Metabolism of fat-soluble vitamins	FALSE
R-HSA-445144	Signal transduction by L1	TRUE
R-HSA-390648	Muscarinic acetylcholine receptors	FALSE
R-HSA-416476	G alpha (q) signalling events	TRUE
R-HSA-189483	Heme degradation	FALSE
R-HSA-8939211	ESR-mediated signaling	TRUE
R-HSA-1482788	Acyl chain remodelling of PC	FALSE
R-HSA-5637812	Signaling by EGFRvIII in Cancer	TRUE
R-HSA-9027307	Biosynthesis of maresin-like SPMs	FALSE
R-HSA-5218920	VEGFR2 mediated vascular permeability	TRUE

Table 3.12: Summary of the NCS scores, MOA, target gene symbols and targets in network of the filtered 11 drugs in the EGFR-TKIR drug-target network.

Drug	NCS	MOA	Targets	Targets In Network
erlotinib	1.78	EGFR inhibitor	EGF; EGFR; GAK; MAP3K19; NR1I2; SLK; STK10	EGF; EGFR
PLX-4720	-1.78	RAF inhibitor	BRAF; FGR; KDR; MAP2K5; MLTK; PTK6; SRMS	BRAF; KDR
hispidin	-1.75	PKC inhibitor	PREP; PRKCB	PRKCB
AG-957	1.72	Protein tyrosine kinase inhibitor	ABL1; EGFR	EGFR
mepacrine	1.63	Cytokine production inhibitor; NFkB pathway inhibitor; TP53 activator	AKT1; MTOR; NFKB1; PLA2G1B; PLA2G2A; PLA2G2D; PLA2G4A; PLA2G6; PLCL1; TP53	AKT1; MTOR
vemurafenib	-1.60	RAF inhibitor	BRAF; CYP3A4; CYP3A5; RAF1	BRAF; RAF1
sirolimus	1.60	MTOR inhibitor	CCR5; FGF2; FKBP1A; MTOR	FGF2; MTOR
atorvastatin	1.59	HMGCR inhibitor	AHR; APOA1; APOB; APOE; CCL2; CD40LG; CETP; CRP; CYP3A5; DPP4; FASLG; HMGCR; IL6; LDLR; LPL; MTTP; PON1; RHOA; SERPINE1; SLCO1B1; VCAM1	IL6
GSK-3-inhibitor-IX	-1.59	Glycogen synthase kinase inhibitor; Lipoxygenase inhibitor	ALOX5; GSK3A; GSK3B	GSK3B
oxindole-I	-1.58	VEGFR inhibitor	AKT1; KDR; RET	AKT1; KDR
gefitinib	1.57	EGFR inhibitor	ABCG2; EGFR; EPHA6; ERBB3; ERBB4; GAK; IFI27L2; IRAK1; MAPK3; VEGFA	EGFR; ERBB3; MAPK3; VEGFA

(GLUT4) for glucose transport (part of vesicle-mediated transport); and signaling by interleukins (part of immune system). These pathways are targeted by the 11 well-characterized LADs, many of them are known to be related to human longevity, the new ones can also be assumed to be new potential pathways that are related to human longevity. The targets of LAD11 and DrugAge drugs were significantly globally mapped to gene expression (transcription) (descendant pathways are related to regulation of TP53 activity) and vesicle-mediated transport compared to a broader targets of MOA drugs, suggesting that the genes/proteins regulating gene transcription are the main target site for longevity-promoting drugs design.

Although many pharmaceutical agents have been identified that may extend lifespan, the scope of available compounds eliciting life extending properties as well as the pathways involved are currently unknown. In this study, databases of GEPs from cells treated with compounds coupled with GESS approaches were leveraged to determine the scope of available compounds that mimic lifespan extending GEPs and may be repurposed for the extension of active living. From a more focused search results that were generated by querying LINCS with the GESs of LAD11, compounds of wortmannin, cinnarizine, SU-4312, arctigenin, dihydroergocristine, lonidamine, quercetin, geldanamycin, CGP-52411, and staurosporine were prioritized from the vote strategy. They share the same MOAs and pathways with LAD11 and also exist in DrugAge database, indicating their ability to elicit GESs similar to GESs of known high-confident LADs and whose mechanisms of action and target pathways are promising to be related to longevity. They can be very promising drugs in extending lifespans for drug repurposing. Furthermore, wortmannin, geldanamycin, lonidamine, dihydroergocristine, SU-4312, staurosporine and quercetin are

also highly ranked from other GESS result summary approaches, which also identify additional LAD candidates of caffeine and pyrazolanthrone, or the DT network of signaling by VEGF pathway. The prioritized drug list for LAD11 also contains many drugs with therapeutic efficacy and unknown drugs or small molecules as novel findings that worth testing. Users can choose from these drug lists according to their research interest. For example, they can choose from the PDs from sirolimus query if sirolimus is more trusted in its effects of promoting longevity. The candidate LADs can be tested in the following experimental studies. For example, they can be first tested in mouse fibroblasts and then in living mice with different dosages for their efficacy in extending life span. Wortmannin is a selective inhibitor of PI3K and can be used to increase the median and maximal lifespan of the fruit fly [132]. Geldanamycin inhibits HSP90, a molecule that plays important roles in the regulation of cell cycle, cell growth, cell survival and apoptosis [128, 65, 157]. KU-0063794 is a small molecule filtered from the DT network of signaling of VEGF pathway. It is an mTOR inhibitor, which is involved in a well known pathway related to longevity [60]. In addition to the identification of new lead compounds, the GESS results may provide additional information about compound characteristics that may be optimized for in vivo therapeutics. Geldanamycin, for example, was identified as a significant lead compound in the GESS results yet causes hepatotoxicity when administered in vivo [66]. Tanespimycin, however, is a structural analogue of geldanamycin and was developed to overcome liver toxicity issues [217].

To date, there are several well established pathways related to longevity yet the scope of LAPs is not well characterized [124, 137]. Results in Figure 3.6B, 3.6C, 3.7 illustrate

the prioritized enriched Reactome pathways performed on molecular targets of compounds with similar GESs to known LADs. These results heavily overlap with known LAPs and may, in part, partially redefine the scope of pathways that may be targeted pharmacologically in order to extend lifespan. Pathways in these results that are also enriched from genes in GeneAge database include PI3K/AKT related pathways, signaling by VEGF, regulation of PTEN gene transcription, regulation of TP53 degradation, insulin receptor signaling cascade, immune system related pathways (signaling by interleukins), SUMOylation of intracellular receptors, and extra-nuclear estrogen signaling from sirolimus query. Pathways of VEGFA-VEGFR2 pathway, signaling by ERBB4, cytochrome P450 - arranged by substrate type, post NMDA receptor activation events, extra-nuclear estrogen signaling, MAP kinase activation, xenobiotics are identified from acarbose query. Pathways of xenobiotics, GPCR ligand binding, neuron system related (Neurotransmitter receptors and postsynaptic signal transmission, activation of NMDA receptors and postsynaptic events, dopamine receptors), SUMOylation of intracellular receptors, extra-nuclear estrogen signaling, intracellular signaling by second messengers, PI3K/AKT signaling in cancer, VEGFA-VEGFR2 pathway, RAS signaling, G alpha (i) signalling events, MAPK family signaling cascades, signaling by RAF1 mutants are supported by multiple query LADs. The direct enrichment on the top 50 PDs from sirolimus query have a very good consistency to the prioritized PPs in sirolimus' FEA results and the prioritized PPs across LAD11 also have a large overlap, making the results more robust and trustworthy. The frequently discovered overall promising LAPs include PI3K/AKT related pathways, intracellular signaling by second messengers, insulin receptor signalling cascade, estrogen and steroid related pathways, neurotransmitter signaling, im-

mune system (signaling by interleukins), signaling pathways involved with VEGF, growth factor receptors (IGF1R, VEGF), TP53, RET, MAPK, RAS, RAF1, AMPK. While many of these pathways are attenuated in people administered LADs and are related to physiological data related to disease prediction (glucose tolerance, hyperlipidemia, hypercholesterolemia), or are associated with cellular phenotypes associated with disease (cell proliferation and inflammation), many putative new LAPs have not been clinically ascribed with disease prevention [84, 25, 73, 151, 106, 27]. Many studies have demonstrated that the downregulation of signaling pathway of IGF-1/PI3K/Akt can extend longevity as well as resistance to oxidative stress in the nematode *Caenorhabditis elegans* [70, 139, 131, 113, 118, 98]. These pathways are promising for human healthy aging and may be pharmacologically targeted to extend lifespan.

The new drugs found by the longevity phenotype query GES from Peters *et al.* (2015) that are highly ranked in GESS result and positively connected with the query GES include BRD-K63954456 (PubChem CID 2202512, a macrophage migration inhibitory factor) and erlotinib. Drugs of pravastatin, amsacrine, thioridazine, exemestane, rosuvastatin, chlorphenamine, gemfibrozil, crizotinib, alprostadil, meloxicam, mepacrine, GW-501516 that have therapeutic efficacy or under FDA study are overlapped between the top 500 ranking drugs in Peters GESS results and the PDs from LAD11 (Table 3.9). It gives more credibility to the overlapped drugs that can be assume to be related to human longevity and healthy aging. The Peters *et al.* (2015) phenotype query also identifies the known LADs or DrugAge drugs of sirolimus, simvastatin, KN-93, phenformin, pyrazolanthrone, arctigenin, caffeine, wortmannin in the top 500 rankings (Figure 3.8B). They can also be added to the candidate

drug list if users are more interested in the human longevity phenotype query. Peters *et al.* (2015) phenotype query also identifies a list of KEGG and Reactome pathways that might be related to human longevity. The KEGG pathways include EGFR-TKIR, ErbB signaling pathway, and glioma. The Reactome pathways that are also identified in the top rankings from LAD11 queries include diseases of signal transduction by growth factor receptors and second messengers, PI3K/AKT signaling in cancer, and neurotransmitter receptors and postsynaptic signal transmission (Table S15). The DT network of EGFR-TKIR pathway also filters a list drugs and target proteins/genes that are involved and can be hypothesized to be related to human longevity. The drugs include sirolimus, erlotinib, PLX-4720, hispidin, AG-957, mepacrine, vemurafenib, atorvastatin, GSK-3-inhibitor-IX, oxindole-I, and gefitinib. The genes include EGFR, MTOR, RAF1, and MAPK3 [159, 175, 207]. These genes can be the targets for designing drug inhibitors to modulate the pathways and extend lifespan.

Most molecular pathways are made up of multiple functional steps, each of which can be subjected to regulations. Therefore, it is possible to identify drugs with different molecular targets within a specific pathway. These drugs will have a similar therapeutic effect. For example, sirolimus/rapamycin, and KU0063/KU-0063794 both inhibit the mTOR pathway. While sirolimus allosterically binds to TOR complex 1 (TORC1), KU-0063794 can suppress the activity of both TORC1 and TORC2 [60]. In addition, many other compounds are under development that can bind to TORs ATP binding site. They can thus inhibit both TORC1 and TORC2 but also target PI3K [11]. Although both TORC1 and TORC2 participate in the mTOR pathway, they have slightly different effects and inhibiting both

of them with ATP competitive inhibitors may prevent tumor proliferation and have an additional antineoplastic effect.

There is a large overlap between putative LADs and antineoplastic drugs, which characteristically have a large number of side effects. Arguably, however, the ability of a drug to act as both a LAD and an antineoplastic agent may in part be explained by its mechanism of action overlap with a longevity pathway. For example, drugs such as rapamycin, that promote cell senescence without directly impairing DNA replication, may have dual effects as LADs and chemotherapeutic agents [16]. However, chemotherapeutic drugs developed as cytotoxic agents or that directly impact DNA replication (gemcitabine) may elicit a greater number of side effects, and have no effect or reduce lifespan in healthy individuals. Another possibility is the level of reversibility a drug imparts on cellular proliferation. While rapamycin drives the conversion from cellular quiescent to senescent states, it preserves the ability of a cell to reenter a proliferative state, other chemotherapeutic agents that are not also LADs may not preserve the ability of cellular re-entry into a proliferative state [16].

Established LADs have diverse mechanisms whereby they regulate organism physiology. While some may directly target cellular receptors (statins), others may indirectly affect cellular function via secondary paracrine or autocrine signaling or indirectly increase longevity via promoting a confounding calorie restriction phenotype (acarbose) [90, 177]. Although some compounds may lack molecular targets in a specific cell type, they may have a direct impact on gene expression that correlates to a longevity GES.

In addition, some pharmaceutical agents may elicit off-target Pleiotropic effects in addition to their known molecular target. Although pathway analyses illustrated in Figure

3.6B and 3.6C were based on known drug targets, the selection criteria for inclusion in the pathway analysis was based on similarity between gene expression profiles and results indicated a high likelihood of recalling drugs with similar MOAs (Figure 3.2). Taken together, the results illustrated in Figure 3.4 and Figure 3.6 identified new potential compounds and pharmacologically targetable pathways that may be involved in longevity as well as providing important direction for pharmaceutical re-purposing and development.

3.4 Materials and Methods

3.4.1 Gene Expression Signature Searching

Gene expression signature (GES) searching (GESS) was performed in R using the *signatureSearch* R/Bioconductor package as previously described by Duan *et al.* 2020 [45]. Among the GESS methods implemented in the *signatureSearch*, *SPsub* conducts GESS correlation-based method (Spearman correlation) to quantify the similarity of a query GES to the GESs in a reference database. Subramanian *et al.* (2017) introduced a gene set-based GESS algorithm (here referred to LINCS method) that uses a bi-directional weighted Kolmogorov-Smirnov enrichment statistic to compute weighted connectivity scores as comparable similarity measure for GESs [186]. In this project, the *SPsub* GESS method was used for systematically testing the recall performance of LINCS drug GESs for known MOA and longevity classifications. For these tests, the query GESs were drawn from LINCS and defined as the top 150 up- and down- regulated genes with expression values for each drug perturbation. The LINCS method was used in the more focused LAD discovery study on GESs of LAD11 as well as one longevity phenotype query GES. Since the query GESs were

obtained from different technologies and organisms instead of just drawing from the LINCS database, the set-based LINCS GESS method was performed. It is well suited in this situation since as a set-based GESS algorithm, it only requires gene sets as input, which are technology agnostic. The query GESs for the LINCS GESS method in this project were the 150 up- and down- regulated gene identifiers, the normalized connectivity score (NCS) was used as a measure of the query GES similarity to GES entries in the LINCS database.

3.4.2 Functional Enrichment Analysis

Functional Enrichment Analysis (FEA) was used to functionally interpret the GESS results as previously described [45]. Specifically, the top 100 unique ranking drug set in the GESS results by $|NCS|$ scores were converted into target gene/protein set and the modified hypergeometric test with duplication support (*dup_hyperG* method) was performed on the target set with duplications to get the enriched functional categories based on the chosen annotation systems, such as Gene Ontology terms in Molecular Function, Biological Process or KEGG, Reactome pathways.

3.4.3 Eleven LADs Selection

The 11 LADs (LAD11) were chosen based on their presence in both compounds in testing (CIT) in longevity assays from the Interventions Testing Program (ITP) at the National Institute on Aging [206], the LINCS and DrugAge databases. In addition, some expert choices were included. The commonly present drugs are: sirolimus, aspirin, simvastatin, resveratrol, curcumin, acarbose, metformin, minocycline. The 3 other well-known drugs from expert choices that can be used to extend lifespan in model organs are tanespimycin,

alpha-estradiol and beta-estradiol. Fuentealba *et al.* (2019) identifies tanespimycin as an anti-ageing drug and experimentally validates in *Caenorhabditis elegans* [58]. The 17-alpha-estradiol is tested in male mice and demonstrated that it can robustly extend both median and maximal lifespan [184]. Beta-estradiol is epimer of the alpha-estradiol. It is tested in *C. elegans* by Ye *et al.* (2014) and identified to increase longevity and resistance to oxidative stress in *C. elegans* [225]. Table 3.1 and 3.4 list the detailed annotation information of the selected LAD11 and their GES sample information including the cell types.

3.4.4 MOA Annotation and Size Cutoff

The full MOA annotations were obtained from the CLUE website at <https://clue.io>. It contains 2,384 drugs in 701 MOA categories. The drugs with MOA annotations (here referred as MOA drugs) were filtered by selecting those that have treatments in LINCS expression database resulting in 2,271 drugs in 678 MOAs. They were used for getting the MOA rankings from DrugAge LAD queries and LAD11 queries. The 334 MOAs with 2,050 unique drugs after size 2 cutoff were used for calculating the recall rates. Since the small MOAs with size 1 or 2 tend to be enriched in the top rankings by their connectivity to longevity, the results were further filtered by selecting MOAs that have 5 or more drugs. This resulted in 1,644 unique drugs in 138 MOAs which was used for the plotting and result discussion for both the recall performance analysis (‘Recall Performance of MOA Categories’ result section) and ranking by connectivity to longevity (‘MOAs Connected with Longevity’ result section). The suitability of a size cutoff of 5 was determined by confirming that the remaining MOAs are representative for the full set. The need for this size filter was that frequently MOAs are split into two where one category contains many drugs, while

other only one or two drugs, *e.g.* Glucocorticoid receptor combined with modulator has one member only, but the agonist version has 33 drugs. The result shows that 103/540 filtered out MOAs have overlapping terms with larger size in the final 138 MOA set. Secondly, the number of overlapped drugs in DrugAge or LAD11 before and after filtering was compared. The result shows that there are 86 overlapped drugs between DrugAge and MOA drugs before filtering and 70 overlapped drugs after filtering, indicating only 16 DrugAge drugs were removed by implementing a size 5 cutoff. For the LAD11, 9 of them overlapped with the MOA drugs before filtering, only one drug acarbose was removed after implementing the size 5 cutoff. So removing the MOA categories that have less than 5 drugs will not result in much information loss. Second, the full 678 MOA categories without size cutoff were ranked by their connectivity to longevity, but the small MOAs with size 1 or 2 tend to rank in the very top or bottom positions. It means that the MOA sizes have large impact on their rankings. To eliminate the MOA size bias, the MOAs were ranked from lowest to largest by their size and their original rankings were recorded. The moving average (MA) scores with window size of 50 were calculated on MOA ranking values and plotted in Figure 3.9. The results indicate that the MOA rankings have large fluctuations on MOAs with small sizes. The fluctuations stabilize at MOA size of 5, meaning that the MOA size tend to have little effect on their rankings. So the size 5 cutoff was applied on MOAs when analysing the results.

3.4.5 Recall Performance of MOA Categories

The MOA recall rates were used as a metric to assess the ability of drugs within an MOA to recall each other. The 334 MOAs with size 2 cutoff were used for calculating the

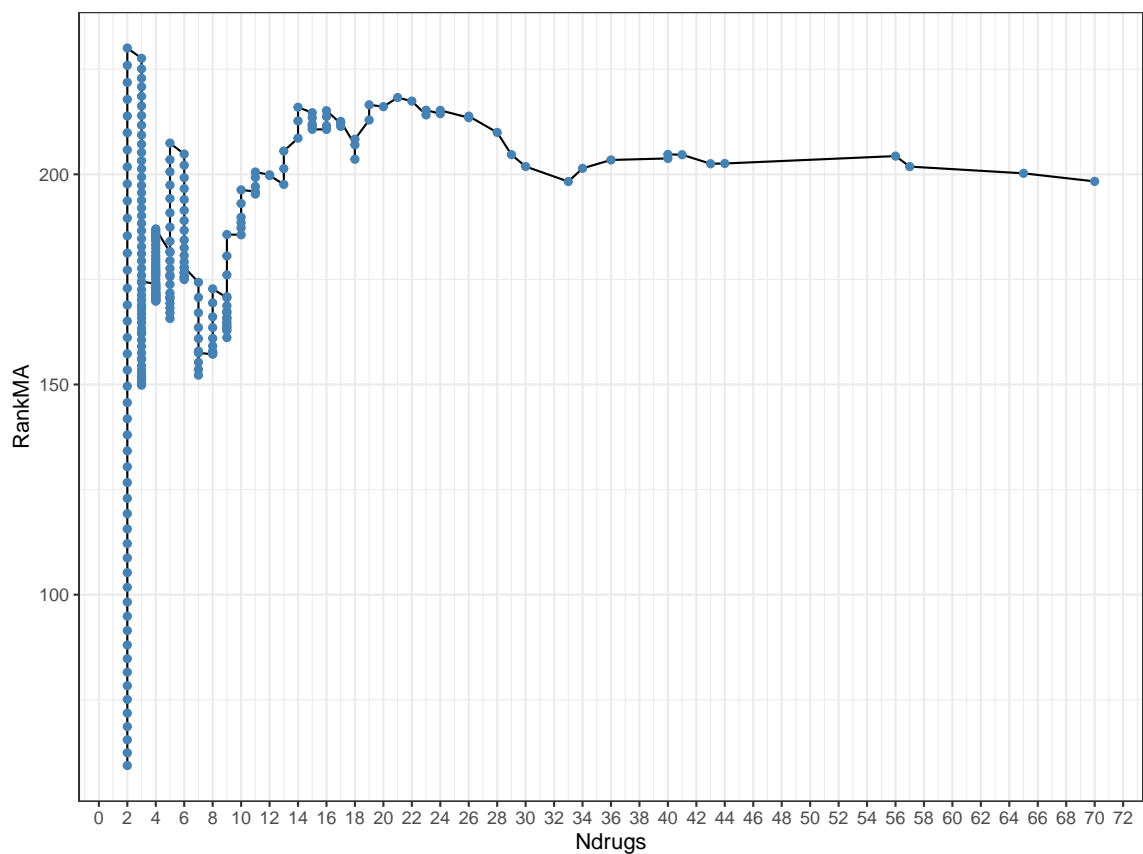


Figure 3.9: Distribution of ranking moving average on MOA sizes using moving window analysis. Ndrugs: number of drugs in MOAs, i.e. MOA size. RankMA: moving average of MOA rankings where MOAs ranked from lowest to largest by their sizes. The moving average of their rankings were calculate with a window size of 50. Since a lot of MOAs have size less than 4 and the RankMA was calculated at each MOA, the dots at small Ndrugs are vertical aligned.

recall rates since it is not possible to calculate meaningful recall rates for MOAs that have only one drug. Then a size 5 cutoff was applied to minimize the MOA size dependency in the MOA performance rankings. First, the GESs of 2,050 drugs in 334 MOAs were queried against the LINCS expression database using the SPsub method. PC3 (prostate tumor) was selected because this cell type has been tested in LINCS with the largest number of compounds. If PC3 cell treatment is not available for a drug, either the MCF7 (breast tumor) cell line or one of the other available cell types were selected. In summary, 2,021 drugs have treatments in PC3 cells, 23 out of the 29 remaining drugs have treatments in MCF7 cells. For the other 6 drugs, one of the available cell types were selected for each one. The MOA recall rates were calculated as illustrated in Figure 3.10. Specifically, drug GESs in one MOA category were iteratively queried against the LINCS expression database with the *SPsub* GESS method. For each drug GES query, CORct scores were appended to the SPsub GESS result, which is the correlation coefficient scores summarized across cell lines by choosing the 67% or 33% quantile of the scores that have larger absolute value. It is similar as the NCSct scores from *LINCS* method. The GESS results from each GES queries in one MOA category were combined into one table. The column-wide median of the absolute CORct scores for drugs in the query MOA across it drug GES queries were calculated as ColMedianCORct and plotted as box plots in Figure 3.2. All MOA categories were ranked from largest to lowest by median of absolute CORct of their drugs across drug GES queries of the query MOA. The rank percentile of the query MOA was set as its recall rate. The lower the MOA recall rate, the better the recall performance of the MOA category.

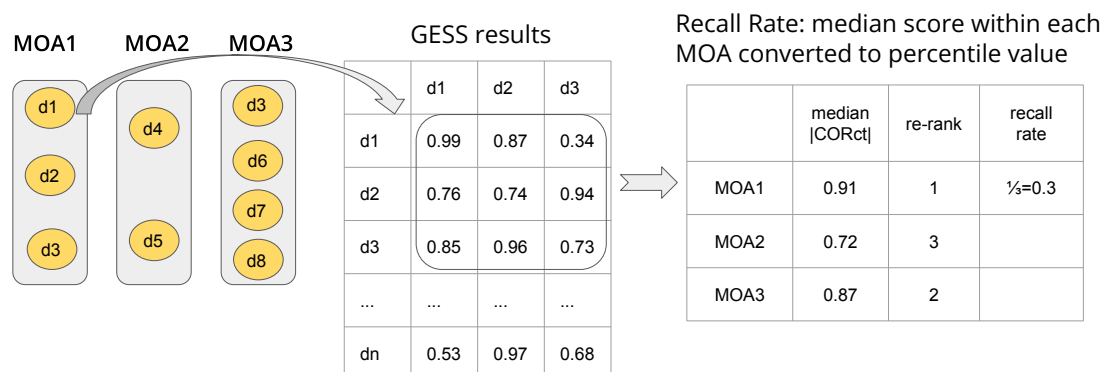


Figure 3.10: Illustration of MOA recall method. Drug GESSs of each MOA category were iteratively queried against the LINCS expression database with the *SPsub* method. All MOAs were ranked from largest to lowest by median absolute CORct of their drugs across drug GESS queries of the query MOA. The rank percentile of the query MOA was set as recall rate.

3.4.6 MOAs Connected with Longevity

To rank MOA categories by their connectivity to longevity, the GESSs of known LADs in the DrugAge [6] database were used to iteratively query the LINCS database with *SPsub* GESS method. There are a total of 112 LADs that exist in both the DrugAge and LINCS databases by matching their PubChem CIDs. Of the 112, 87 LADs (LAD87) have GESSs in normal cells in LINCS database. Since the LAD87 are tested in different numbers of normal cell lines in LINCS, it results in a total number of 495 query GESSs for LAD87 in 12 normal cells, instead of the multiplication of the drug and cell number. The latter were used to query LINCS with the *SPsub* method. Similar to the recall performance analysis, the *SPsub* GESS results from each of the LAD GESS queries were cell type summarized by CORct scores. The GESS results from all of the GESS queries were then combined into one table. The MOA categories were ranked from highest to lowest by median of absolute CORct of their drugs across all 495 GESS queries. The full MOA categories were used for the

calculation, while applying a size cutoff of 5 to minimize the MOA size bias. The same MOA ranking approach was also applied to the GESs of LAD11. The consistency of their MOA rankings was compared by counting the number of overlapped MOAs in their top 20s as well as calculating the Spearman correlation coefficient of the global rankings. The MOA recall rates obtained from the above ‘Recall Performance of MOA Categories’ section were also appended to the result table as reference of their recall performance. The top ranking MOAs were considered to be associated (connected) with human longevity.

3.4.7 Drug-Target Pathway Map Plotting

The targets of the LAD11 set were mapped to Reactome pathways and visualized with firework plots. Since a local implementation does not exist, the fireworks plotting had to be performed with Reactome’s [62] online service. The highest level human Reactome pathways that are enriched in the target proteins of LAD11 were highlighted as colored branches in the firework plot manually. To contrast the enrichment differences, the Reactome enrichment results on each of the three drug sets (LAD11, DrugAge drugs and MOA drugs) were obtained by the hypergeometric test on their unique target set. The top level Reactome pathways in their enrichment results were subsetted and compared in a heatmap plot by combining the enrichment results into one table. The adjusted p-values, normalized counts, number of target genes in the test gene set and pathway size were also plotted in the heatmap.

3.4.8 Permutation Tests

The permutation test was used to compute a Rank Robustness Score (RRS) for both the GESS and FEA results to test the robustness of the rankings obtained from the

GESS and FEA results. The RRS was calculated for 5, 10 and 15 percent randomization. As a result, RRS values decreased in the corresponding order. When the query GES consists of up and down-regulated gene sets (*e.g.* *LINCS* method) permutation testing involves randomly sampling 5%, 10% and 15% of the query genes and replacing them with other genes in the reference database. When the query GES consists of expression intensity values of up and down subsetting genes for the *SPsub* method, permutation testing also involves randomly sampling 5% , 10% and 15% of the query genes and replacing them with expression values of other genes drawn from the reference database. The randomization was repeated 100 or 1000 times. Then the GESS methods were applied to both the original query and randomized queries. The *dup_hyperG* method was used as the FEA method, where the targets of the top 100 unique drugs in a ranked GESS result served as test set. The drug rankings in the permutation GESS results were compared to their rankings in the original result. The RRS values were expressed as the percent of the permutation times that the drug rankings in the permutation result are within the +2 and -2 ranking windows of the rankings in the non randomized GESS result. The same applies to pathways in the FEA results. Figure 3.11 illustrates the processes of permutation testing.

3.4.9 Drug Prioritization Strategy

The GESS result from each query GES contains multiple NCS scores for LINCS compounds across different cell types. In order to get the cell type summarized score for each compounds, the GESS result can be summarized by three metrics: (1) NCSet scores. They are the quantile based cell type summarized NCS scores obtained from the LINCS method. The compounds' NCS scores were summarized across cell lines by choosing the

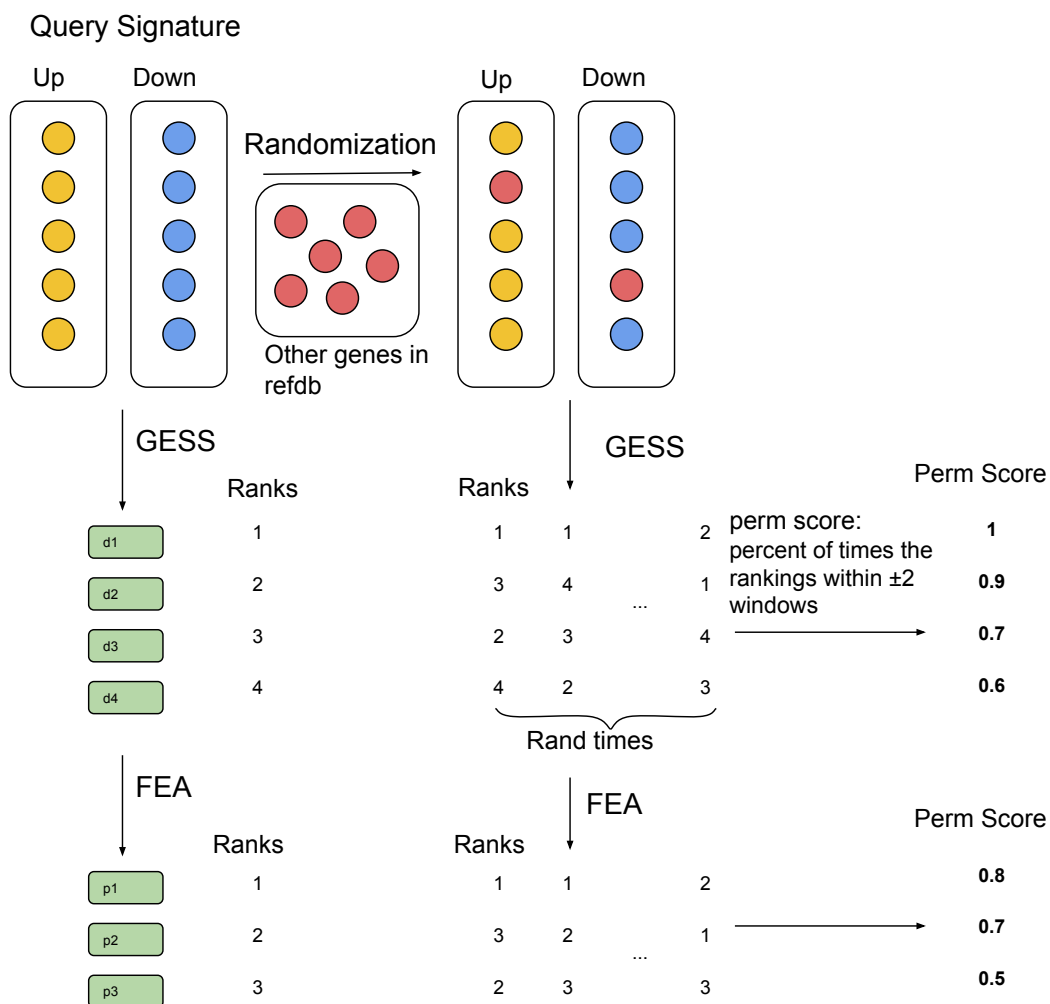


Figure 3.11: Permutation testing method. Query GESs were randomized at 5% , 10% or 15% of the GES query genes by iteratively replaced with randomly selected genes from the reference database 1000 times prior to performing the GESS/FEA analysis. The permutation results were compared to the original result rankings and the permutation/robustness scores were calculated as the percent of permutation times that the drug/pathway rankings of the permutation results within the +2 and -2 ranking windows of the non randomized query GESS results.

67% or 33% quantile of the scores that have larger absolute values; (2) NCSunique scores. They are obtained by choosing the compounds' largest NCS scores across cell types. It was done by ranking the full GESS results from largest to lowest by NCS scores. The GESS result table was then uniquified by the compound name column. (3) NCStissue scores. As illustrated in Figure 3.12A, it first chose the compounds' NCS score in primary/normal cell type that match the query cell type/tissue. If the compound has no NCS score in the matched primary cell, the NCS score in the immortal/tumor cell line of the query tissue was chosen for that compound. If the compound is tested in cell types that do not match either the primary nor the immortal cell line, the largest NCS score tested in other cell types was chosen. The GESS results were then reduced to the compound level by including one cell-type summarized NCS score for each compound.

The compounds in GESS results from multiple GES queries of one query LAD (*e.g.* GESs of the query LAD treated in different cell types, from different technologies or organisms) can be summarized by several methods including voting strategy, summarized NCS scores (cell type quantile or maximum NCS scores) across multiple queries (mqNCSct, mqNCSunique) or just using the compound rankings from one specific GES query. The voting strategy uses frequency of how often compounds showed up at top of each individual GESS result where the compounds in each GESS result can be ranked by the above introduced three metrics. The voting strategy on different cell type summarized NCS scores results in three voting methods (voteNCSunique, voteNCStissue, voteNCSct). As illustrated in Figure 3.12B, the cell type summarized GESS results from different GESs of one query LAD were combined into one table. The rank-transformed table was turned into TRUE

(1) or FALSE (0) values by setting the rank cutoff (e.g. 400 in this project). The LINC compounds in GESS results were then summarized across queries by ranking from largest to lowest by row sums of cell type summarized NCS scores after cutoff. The re-ranked compounds in GESS results can then be stratified/classified into 3 layers (Figure 3.12C). Layer 1 contains drugs sharing MOAs with the combined MOA set of the 11 query LADs. Layer 2 contains drugs targeting the same Reactome pathways as the query LADs. The other compounds were stratified into layer 3. Compounds in each layer remain the rankings summarized from voting strategy. The top ranking compounds in each layer (here 15 in layer 1 and 2, 20 in layer 3) were selected as the final prioritized drugs (PDs). Drugs in layer 1 and layer 2 with known annotations can be hypothesized for drug repurposing. Layer 3 contains many unknown compounds and small molecules, which can be redeemed as new findings of unknown small molecules that worth testing. Their corresponding cell type summarized NCS scores across GESSs of the query LAD and annotations were plotted in heatmaps. The top 50 PDs for each LAD11 were then summarized by ranking the drugs from largest to lowest by the number of times that they were identified/supported by the query LAD11 (Nsupport).

3.4.10 Voting Strategy on FEA Results

The Reactome pathways in FEA results from GESSs of each query LAD were also summarized with voting strategy in the similar way as the GESS results to get the prioritized Reactome pathways (PPs) for each query LAD. Each of the FEA results from the in vitro GESSs of the query LAD was ranked from lowest to largest by the adjusted p-values. The pathway rankings transformed from the adjusted p-values for different GES queries were

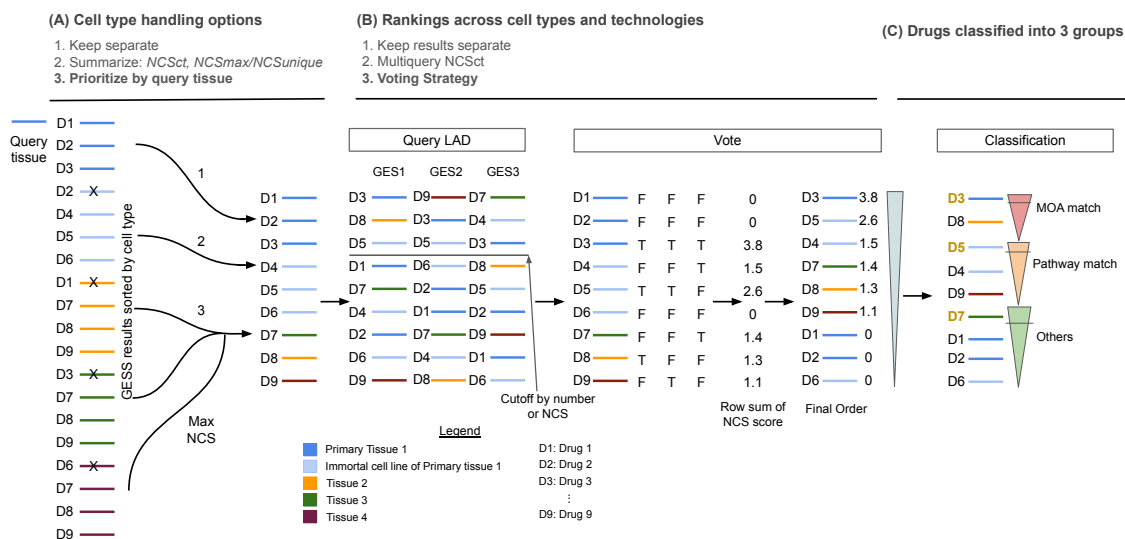


Figure 3.12: Illustration of the NCS tissue score (A), voting strategy (B) and classification (C). The compounds' NCS scores in cell lines that matches the query tissue (primary cell has first priority and then immortal cell line) are selected. The NCS scores in other cell lines are ignored. The maximum NCS scores are selected for compounds that do not have treatments in the matched cell lines. The cell type summarized GESS results from different GESSs of one query LAD are combined into one table. The rank-transformed table is turned into TRUE or FALSE values by setting the rank cutoff (e.g. 400). The compounds in GESS results are then summarized across queries by ranking from largest to lowest by row sums of cell type summarized NCS scores after cutoff. The drug rankings from voting strategy are then classified into three layers (layer 1: drugs matching MOAs with query LADs; layer 2: drugs matching target pathways with query LADs; layer 3: others). Drugs in each layer remain the original voting order. The top ranking drugs in each layer are selected as the final prioritized drugs.

combined into one table and the rank cutoff was applied (*e.g.* 100 in this project). The Reactome pathways in FEA results were then summarized across queries by ranking from lowest to largest by row means of rankings after cutoff multiplied with a size correction factor. The latter was calculated by multiplying the total number of GES queries divided by the number of supported queries after rank cutoff to give it a larger value to pathways that are supported by fewer queries to make it rank lower, thus favor the pathways that are supported by more queries (Figure 3.13). To visualize the result, the top 50 PPs were plotted in form of a heatmap. The color key indicates the rankings. To indicate which of the pathways are associated with longevity, the LAP annotation was obtained by applying hypergeometric test on the longevity associated genes (LAGs) from the [GeneAge](#) database. The enriched functional categories were filtered with the adjusted P-value cutoff of 0.05. The top 50 PPs for each of the LAD11 were then summarized by ranking the pathways from largest to lowest by the number of their query LADs support (Nsupport). The pathways with more than or equal to 3 LAD supports were plotted in a heatmap.

3.5 Supplementary Material

3.5.1 Supplementary Methods

GESS Score Distributions

The GESS score (*e.g.* NCS) distributions in the GESS results from drug or phenotype GES queries were subsetted and plotted as histograms, which show the early enrichment NCS scores (less than -1 or greater than +1) after setting count cutoff as 100 to better display and compare the NCS distributions at left and right extremes. The P-values of the

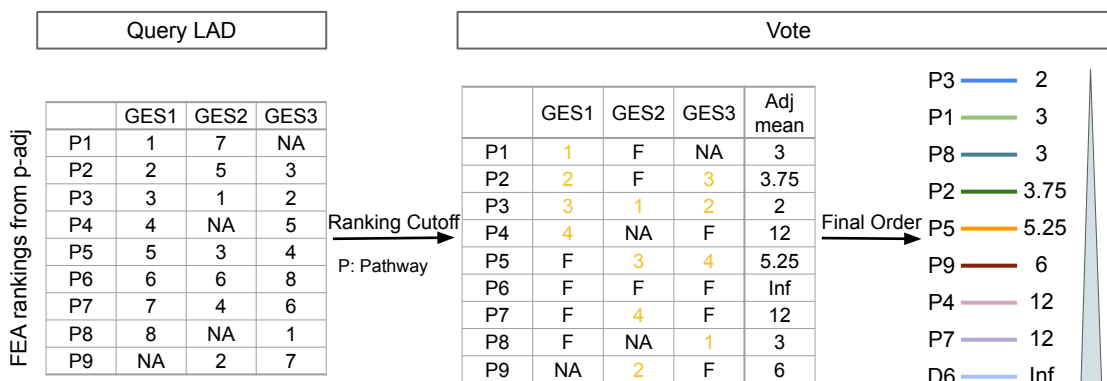


Figure 3.13: Voting strategy on FEA results. Each of the FEA results from the GESs of one query LAD was ranked from lowest to largest by the adjusted p-values and combined into one table. The rank cutoff of 100 was applied. The pathways in FEA results were summarized across queries by ranking from lowest to largest by row means of rankings after cutoff multiplied with a size correction factor. The latter was calculated by multiplying the total number of GES queries divided by the number of supported queries after rank cutoff to give it a larger value to pathways that are supported by fewer queries to make it rank lower, thus favoring the pathways that are supported by more queries.

weighted connectivity score (WTCS) from the LINCS algorithm were displayed as color key.

To compare the NCS distribution from a drug GES to a random GES query, one random query GES was generated by randomly sampling of 150 up and down-regulated genes present in the LINCS database as negative control. The first ranking drug term in GESS result from itself query was removed to avoid giving advantages to the drug queries whose GESs were drawn from the reference database.

Hierarchical Clustering of 11 LADs

The LAD11 were clustered by hierarchical clustering on Spearman correlation coefficients ($SPcor$) of their GESS results rankings. Specifically, the GESS results were ranked from largest to lowest by $|NCS|$ scores after filtering out zeros. The $SPcor$ were calculated on GESS rankings pairwise across LAD11. The $1 - SPcor$ was used as distance metrics for

clustering, where the clustering method was set as `complete`. Since different queries have different drug cell terms in GESS results after zero filtering, the correlation between each pair of variables was computed using all complete pairs of observations on those variables. It was done by setting the `use` argument of the `cor` function as `pairwise.complete.obs`. The overlapped Reactome pathway for LADs in each cluster was obtained by mapping Reactome pathways for each LAD with `dup_hyperG` method from the `signatureSearch` package on their targets and the intersected Reactome pathways across LADs in each cluster were obtained as the common pathways of each cluster.

Test of Drug Prioritization Methods

The drug prioritization methods (DPMs) described under ‘Methods and Materials’ section include `voteNCSunique`, `voteNCStissue`, `voteNCSct`, `mqNCSct`, `mqNCSunique` and compounds ranked by `NCSunique`, `NCStissue`, `NCSct` scores on one selected query sample. The DPMs are first tested on in vitro samples (GESs drawn from LINCS database in different human cell lines from L1000 technology) and in vivo samples (GESs from differential expression analysis on samples of drug treatment in different mouse tissues from RNA-Seq or Microarray technologies) for the 9 out of 11 LADs that have MOA annotations. Thus, each query LAD has multiple query samples/GESs due to different cell types and technologies. The DPMs can be used to get the summarized drug ranking list across multiple GESs for each query LAD. The drugs can be ranked based on only in vitro samples or on all samples for the voting strategy to compare the performance difference of including only LINCS samples or all samples to make a decision on which samples to use. Here only `NCSct` score are used as cell type summarized score, resulting in methods of `voteNCSct-LINCS` on in

vitro/LINCS samples or voteNCSct-all on all samples. The mqNCSct methods were also applied on LINCS samples or all samples for 9 LADs to determine which type of samples is better to be ranked on. The DPM performance is measured by the number of drugs sharing the same MOAs as the query LAD that each methods can fish up/retrieve in their top 100 ranking drugs (Nretrieve). In other words, the methods are compared by showing which method could retrieve more drugs sharing MOAs as the query LAD in their top 100s.

The DPM performance was further compared by applying on a larger number of query drugs (57 in this case including the 9 LADs) and the methods were ranked from largest to lowest by mean of the Nretrieve across 57 query drugs. The query drug set was determined by selecting 5 drugs in each of the top 10 ranking MOAs by recall performance that have at least 10 drugs in it. The selected drugs should also have at least 3 treatments in normal cells in LINCS database. It results in 150 GES queries drawn from LINCS database for 50 drugs, each have 3 normal cell treatments. Finally, 57 drugs including the 9 LADs are selected for the systematic test of the PDMs. Each drug have 3 samples in normal cells, resulting in a total number of 171 query GESs that are iteratively queried against LINCS database with LINCS method.

Drug-Target Network Creation

If one pathway was supported by several LADs queries, the top 100 drugs in the GESS results from each LAD query were combined to build the DT network by depicting the drug-target interactions of the combined drug set to genes/proteins in the selected pathway.

3.5.2 Supplementary Sections

Proof-of-Concept with Three Well-Characterized Drugs

The feasibility of GESS/FEA workflow was tested in advance by querying GESs of WCD3 (vorinostat, alvocidib and chlorpromazine) with permutation tests. It was used as a proof-of-concept that the GESS/FEA workflow using the *LINCS* and *dup_hyperG* algorithms can retrieve similar drugs and target pathways of query drugs. First, the significance of the GESS results from drug GES queries was explored by comparing to a random GES query to demonstrate a substantial difference from random. The distributions of the GESS scores from WCD3 GESs were compared with random queries and the result shows that drug GESs queries have more extreme and significant early enrichment GESS results compared by a random query (Figure S2A). To demonstrate the feasibility of GESS/FEA workflow in retrieving similar drugs and target pathways, the distributions/positions of drugs in the same MOAs as query drugs in their GESS results and pathways directly targeted by the query drug in their FEA results were shown in Figure S2B and C, respectively. The result shows that many ‘correct’ drugs or pathways are enriched in the top ranking positions above the red line in the GESS and FEA results for all of WCD3. Even though some drugs target a wide spread of pathways, many of them matched in the top ranking pathways in FEA results. These results indicate the feasibility of LINCS GESS method in identifying known drugs or unknown compounds that share similar targets/MOAs as the query drugs and the chosen GESS/FEA workflow can discover the correct target pathways of the query drugs.

LAD11 Clustering

The LAD11 was clustered by Spearman correlation of their GESS result rankings (Figure S3A, GESS SP), structural similarity (Figure S3B, Structure), Jaccard index of overlapped targets (Figure S3C, Targets), and enriched GO BP terms (Figure S3D, BP). Estradiol and alpha-estradiol are clustered together by all of the above 4 metrics. Drugs of curcumin, resveratrol, aspirin tend to be clustered together by Targets and Structure similarity (Figure S3E). Drugs of simvastatin and sirolimus, minocycline and tanespimycin tend to be clustered together by GESS SP and Structure similarity (Figure S3F). The tree diagrams comparison results suggest that the most similarity exists between GESS SP and Targets with an entanglement score of 0.25, the next is GESS SP and Structure (0.33), the next is Structure and Targets (0.41), and the last is Targets and BP (0.43).

Drug Prioritization Methods Tests

In order to get prioritized drugs (PDs) in GESS results for each query LAD summarized across different cell types and technologies, several drug prioritization methods (DPMs) are proposed including voting strategy that uses frequency of how often drugs show up at top of each individual GESS result where compounds in GESS result are ranked by cell type summarized NCS scores (NCSunique, NCStissue, NCSct), the rank transformed table is turned into binary table by setting rank cutoff and summarized across queries by row sums of three cell type summarized NCS scores resulting in three vote methods (voteNCSunique, voteNCStissue, voteNCSct). The drugs can also be prioritized by NCSct or NCSunique scores across multiple queries (mqNCSct, mqNCSunique) and compounds rankings simply

based on one selected query sample for each query LAD ranked by NCSunique, NCStissue or NCSct scores. The detail description of the three cell-type summarized scores and DPMs are available at the ‘Materials and Methods’ section. The proposed DPMs are first tested on 9 out of 11 LADs that have MOA annotations (Table S6). The voteNCSct and mqNCSct methods are also applied on only in vitro/LINCS query samples (voteNCSct-LINCS) compared to all (in vitro and in vivo) samples (voteNCSct-all) for the 5 LADs that have both in vitro and in vivo samples to determine which type of query samples are better to be used for ranking. The methods performance is measured by the number of drugs sharing the same MOAs as the query LAD that each methods can fish up/retrieve in their top 100 ranking drugs (Nretrieve). The details of the methods testing is available at the ‘Supplementary Methods’ section. The test result shows that even though different DPMs perform best for different query LADs, overall the voteNCSunique method has the best performance (maximum mean Nretrieve of 3.8). It also shows that summarizing the GESS results on only in vitro query samples is better than from all queries by including in vivo samples. The DPMs were then tested on 57 drugs that were selected from 10 MOAs with top recall performance (Table S7 and Figure S4). The methods rankings on 57 drugs have a very good consistency with the test result on 9 LADs. It further demonstrates that the vote strategy has the best performance (largest mean Nretrieve of 5.0) compared to others. As to the cell type summarized scores for ranking compounds in GESS result, NCSunique is better than NCStissue and NCSct. So, the voteNCSunique method is chosen as the final choice with drug stratification to prioritize drugs in GESS results from LAD11 queries individually for known and novel LADs discovery.

GESS Results Summary Across LAD11

In addition to applying voting strategy on each of the LAD11, the GESS results from all of the in vitro LAD11 GES queries in starred cell types were summarized with row sums of |NCS| scores after applying strict cutoff on individual GESS results to arrive at a combined drug ranking list. Two types of cutoffs were applied, one is filtering the compounds in GESS result that have $RRS5 \geq 0.6$ and $NCS! = 0$ (Figure S8A), the other one is filtering drugs in the GESS result from each LAD query that sharing similar targets as the query LAD and also with $NCS! = 0$ (Figure S8B). The drugs in GESS results across LAD11 GES queries were then ranked from largest to lowest by row sums of |NCS| scores. As the query LADs have different target sites and structures, they fish up different drugs. For example, lonidamine is only identified in the top rankings from the aspirin query. Among the top 50 combined ranking drugs and unknown small molecules, several drugs in DrugAge database are identified. They are lonidamine, geldanamycin, wortmannin and caffeine (a methylxanthine alkaloid, acts primarily as an adenosine receptor antagonist in the central nervous system (CNS) with psychotropic and anti-inflammatory activities. It inhibits the adenosine-mediated downregulation of CNS activity, thus, stimulating the activity of brain. This agent also promotes neurotransmitter release that further stimulates the CNS) (panel A), geldanamycin, wortmannin, pyrazolanthrone (a derivative of anthrone. It is used in biochemical studies as an inhibitor of c-Jun N-terminal kinases (JNKs) [13]), dihydroergocristine and caffeine (panel B). They can also be promising LADs that worth testing. Table S9 shows their MOA and target annotations.

3.5.3 Supplementary Figures

3.5.4 Supplementary Tables

Table S1: Additional annotation information of the 14 query drugs (LAD11 and WCD3) such as up and down-regulated gene sets of query GESs drawn from LINCS database, enriched functional terms (GO MF, BP, Reactome) from their targets with adjusted p-value cutoff of 0.05.

3.6 Data Availability

The source code generated and used by this project is hosted on [GitHub](#). Larger result sets (*e.g.* larger Supplementary tables) are available on Synapse ([10.7303/syn27069757](#)). The full GESS/FEA result tables from LAD11 and WCD3 queries with RRS are available under [syn27074478](#). For example, the GESS and FEA results (Reactome annotation) with RRS for the sirolimus query in SKB cell are shown in tables named as `sirolimus_gess_tb.tsv`, `sirolimus_dup_reactome_tb.tsv`, respectively. The prioritized GESS result tables for LAD11 from VoteNCSunique method are organized under [syn27074560](#). The prioritized Reactome pathway result tables for LAD11 from voting strategy are hosted under [syn27074585](#).

Table S2: Complete table of recall rates and other metrics for the 334 MOA categories with size 2 cutoff.

Table S3: Complete table of MOA rankings with no size cutoff from DrugAge LAD87 queries.

Table S4: Complete table of MOA rankings with no size cutoff from LAD11 queries.

Table S5: Combination table of Reactome enrichment results from the unique target set of LAD11, DrugAge drugs, and MOA drugs with the hypergeometric test for all of the Reactome pathways with top-level pathway information.

Table S6: Performance test of drug prioritization methods on 9 LADs. The numbers next to drug names indicate the number of drugs sharing the same MOAs as the LAD from the drug MOA annotations obtained from CLUE website. The numbers in the table body indicate the number of drugs sharing MOAs as query drug retrieved by different methods in their top 100 rankings (Nretrieve). The value is NA when the query samples are not available for some LADs. The numbers in the parenthesis indicate the number of query samples for the methods used to rank compounds in GESS results. The columns in this table is ordered from largest to lowest by mean of Nretrieve across 9 LADs.

	voteNCSunique	NCSunique	voteNCStissue	mqNCSunique	NCStissue	voteNCSct-LINCS	mqNCSct-LINCS	NCSct	voteNCSct-all	mqNCSct-all
sirolimus 18	14 (5)	9	10 (5)	9 (5)	6	14 (5)	13 (5)	11	11 (18)	7 (18)
metformin 8	1 (2)	1	2 (2)	1 (2)	2	0 (2)	0 (2)	0	0 (12)	0 (12)
resveratrol 22	2 (3)	2	1 (3)	2 (3)	1	2 (3)	2 (3)	1	1 (9)	0 (9)
simvastatin 9	6 (1)	6	3 (1)	6 (1)	3	7 (1)	7 (1)	7	NA (1)	NA (1)
acarbose 4	1 (2)	1	1 (2)	1 (2)	1	0 (2)	0 (2)	0	0 (6)	0 (6)
curcumin 84	2 (1)	2	3 (1)	2 (1)	3	0 (1)	0 (1)	0	NA (1)	NA (1)
aspirin 58	2 (1)	2	6 (1)	2 (1)	6	1 (1)	1 (1)	1	NA (1)	NA (1)
alpha-estradiol 27	3 (1)	3	2 (1)	3 (1)	2	1 (1)	1 (1)	1	NA (1)	NA (1)
estradiol 27	3 (3)	4	1 (3)	1 (3)	2	0 (3)	1 (3)	1	0 (5)	0 (5)

Table S7: Performance test of drug prioritization methods on 57 drugs. The numbers next to drug names indicate the number of drugs sharing the same MOAs as the query drug from the drug MOA annotations obtained from CLUE website. The numbers in the table body indicate the number of drugs sharing MOAs as query drug retrieved by different methods in their top 100 rankings (Nretrieve). The columns in this table are ordered from largest to lowest by mean of Nretrieve across 57 drugs. The numbers next to the NCSunique, NCStissue and NCSct methods indicate the method is applied on sample/cell 1, 2, 3, respectively for the query drug. Their mean numbers are mean of the 3 columns.

Table S8: Combined PDs across LAD11 queries from their individual top 50 PDs. The numbers in the LAD columns indicate layer information.

Table S9: MOA and targets annotation of the identified LADs in Figure S8.

Drug Name	Targets	MOA
caffeine	ADORA1; ADORA2A; ADORA2B; ADORA3; ATM; ...	Adenosine receptor antagonist; Diuretic; Phosphodiesterase inhibitor
lonidamine	CFTR; GCK; HK1	Glucokinase inhibitor
wortmannin	ATM; ATR; MTOR; MYLK; PI4KA; ...	PI3K inhibitor
geldanamycin	HSP90AA1; HSP90AB1; HSP90B1	HSP inhibitor
pyrazolanthrone	LRRK2; MAPK10; MAPK8; MAPK8IP1; MAPK9; ...	JNK inhibitor
dihydroergocristine	ADRA1A; ADRA1B; ADRA1D; ADRA2B; ADRA2C; ...	Adrenergic receptor antagonist; Prolactin inhibitor

Table S10: GO MF, BP, KEGG and Reactome pathway enrichment results from genes in GeneAge database with hypergeometric test after setting adjusted p-value cutoff as 0.05.

Table S11: Combined Reactome pathways across LAD11 queries from their individual top 50 prioritized pathways (PPs). The numbers in the LAD columns indicate that the pathway shows up in the top 50 PPs of the query LAD.

Table S12: Reactome enrichment results on top 50 prioritized drugs from sirolimus. The Reactome pathways are enriched from the *dup_hyperG* method on the target set of drugs with an adjusted p-value cutoff of 0.05.

Table S13: Complete table of Peters GESS result from LINCS method.

Table S14: Complete table of Peters FEA result on KEGG pathways from *dup_hyperG* method.

Table S15: Complete table of Peters FEA result on Reactome pathways from *dup_hyperG* method.

Table S16: The elaborate annotations for all of the prioritized DrugAge drugs in this project including drug description, tested assays, publications, *etc.*

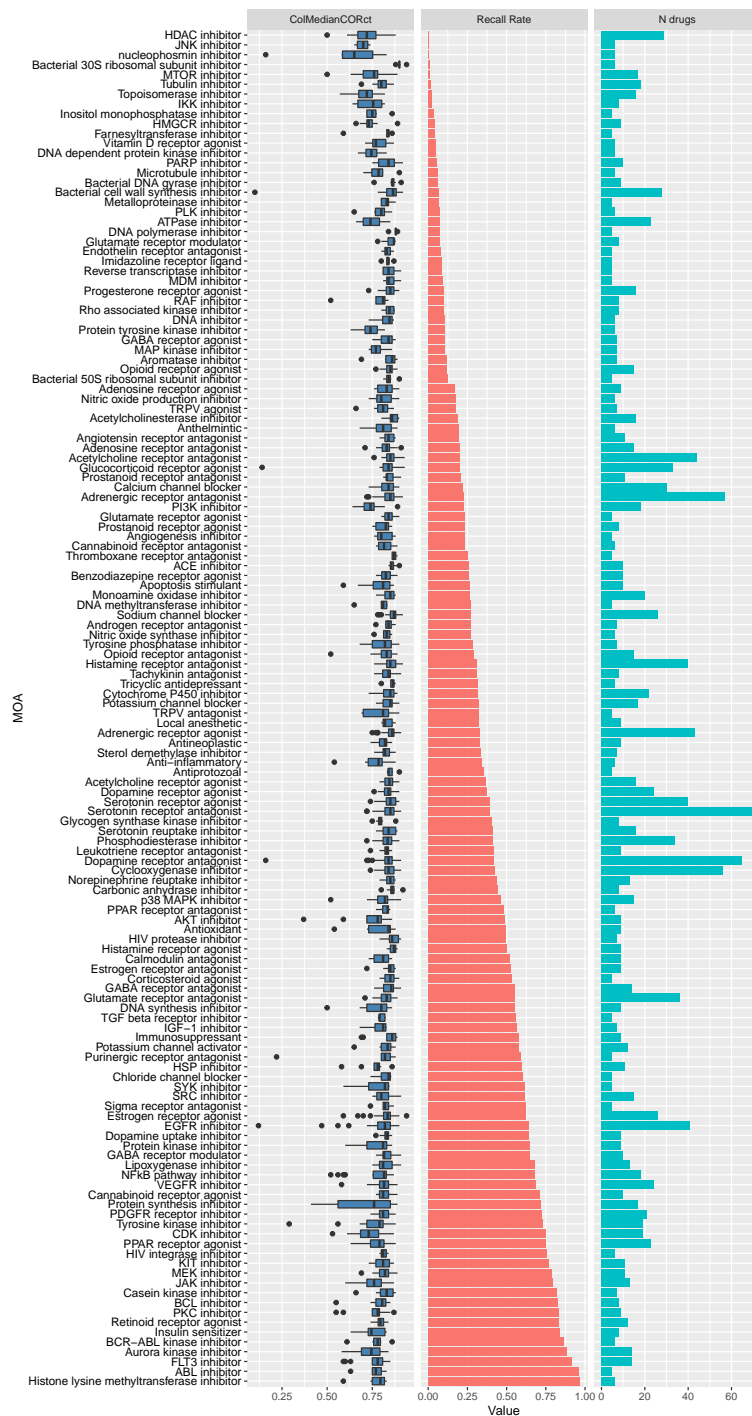


Figure S1: MOA recall rates for all of the 138 MOAs with size 5 cutoff. ColMedianCORct: distribution of column-wise median absolute CORct for MOA drugs from its iterative GES queries. The complete table of recall rates and other metrics for the 334 MOAs with size 2 cutoff is at Table S1.

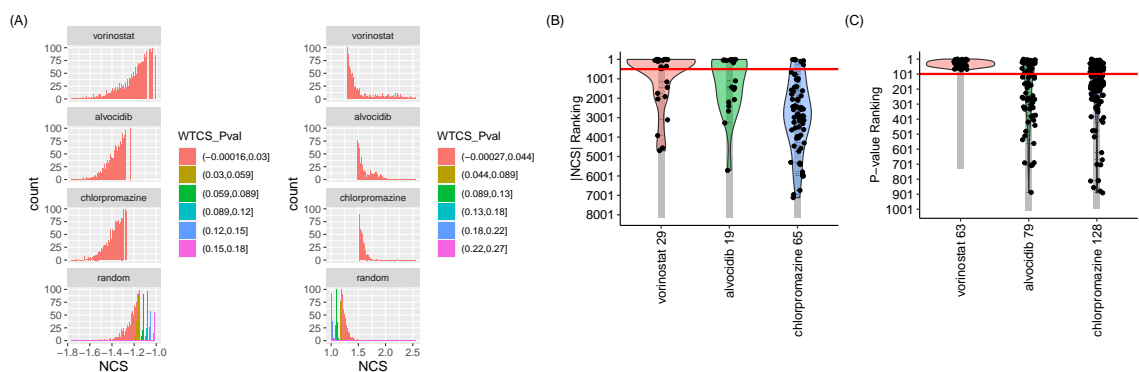


Figure S2: NCS score distributions from WCD3 GES queries compared to random query (A), Rankings of drugs in the same MOAs as query drugs in their GESS results (B) and enriched pathways from target set in DrugBank and CLUE resources of query drugs in their FEA results (C) for the WCD3. The GESS results are ranked from highest to lowest by absolute NCS scores, the FEA results are ranked from lowest to highest by adjusted P-values. The numbers next to the drug names indicate the number of dots. The red line is used as a cutoff showing that the above part should be focused and indicating whether the GESS/FEA results could retrieve similar drugs or pathways in the top rankings.

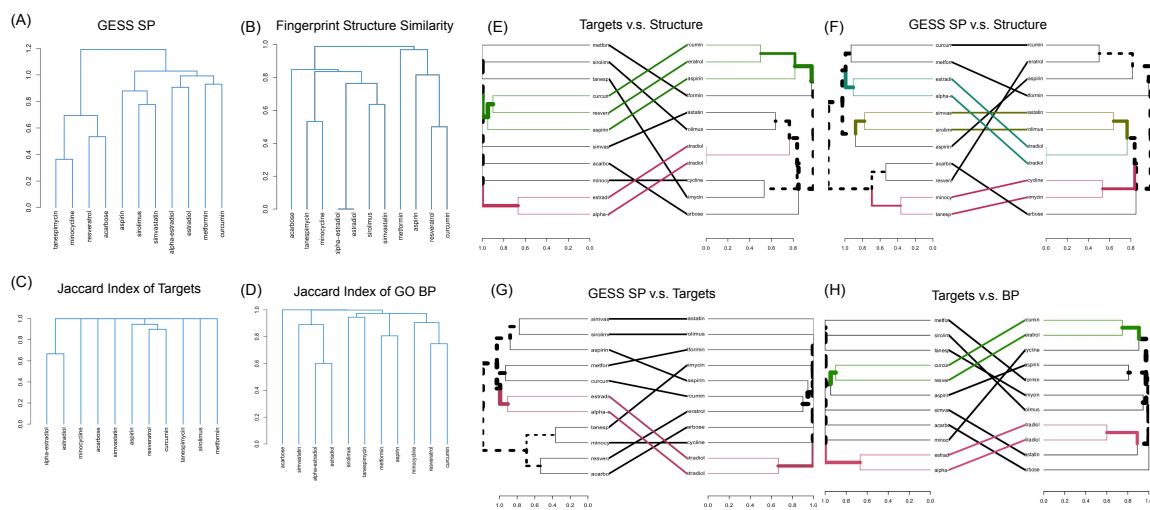


Figure S3: 11 LADs clustered from different approaches. 11 LAD clustered by Spearman correlation of NCS scores in GESS results (A), structural similarity of fingerprints (B), Jaccard index of their targets (C) and enriched GO BP terms (D). (E-H) Comparisons of cluster dendrograms in panels A-D. The Jaccard index was calculated by dividing the number of intersecting targets or enriched functional categories by the number of union terms. The 11 LADs were then clustered by the $1 - \text{Jaccard index}$ as distance. NCS scores equal to zeros were removed prior to performing Spearman correlation analyses (A, F, G). The similarity between the two trees was measured by entanglement score, which is a measure between 0 (no entanglement) and 1 (full entanglement), lower values mean better similarity. The entanglement scores for E-H are 0.41, 0.33, 0.25, 0.43, respectively.

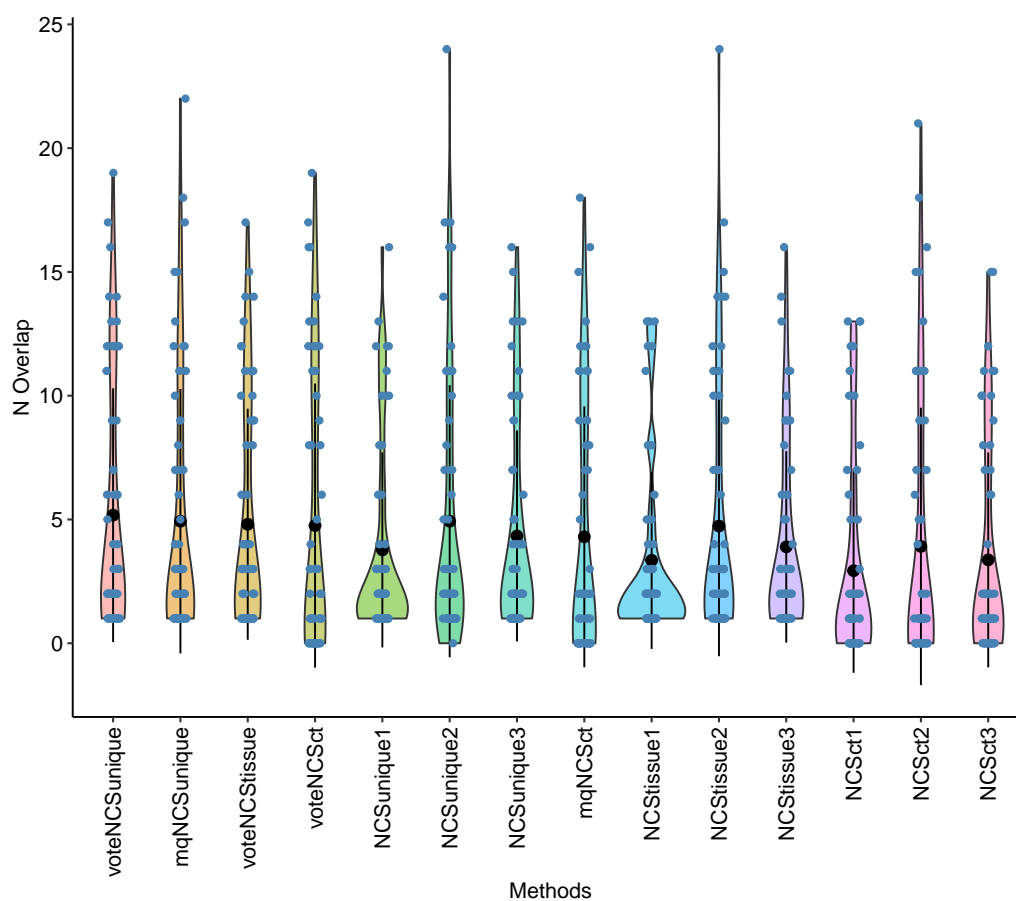


Figure S4: Violin plot of the test result of drug prioritization methods on 57 drugs. The blue dots indicate the number of retrieved drugs sharing MOAs as query drug by methods in their top 100 rankings (N Overlap). The black dot is mean of the numbers. The methods are ordered from largest to lowest by mean values. The numbers next to the NCSunique, NCStissue and NCSct methods indicate the method is applied on sample/cell 1, 2, 3, respectively for the drug query. their mean value is calculated by mean of the numbers across 3 samples. The corresponding table is at Table S7

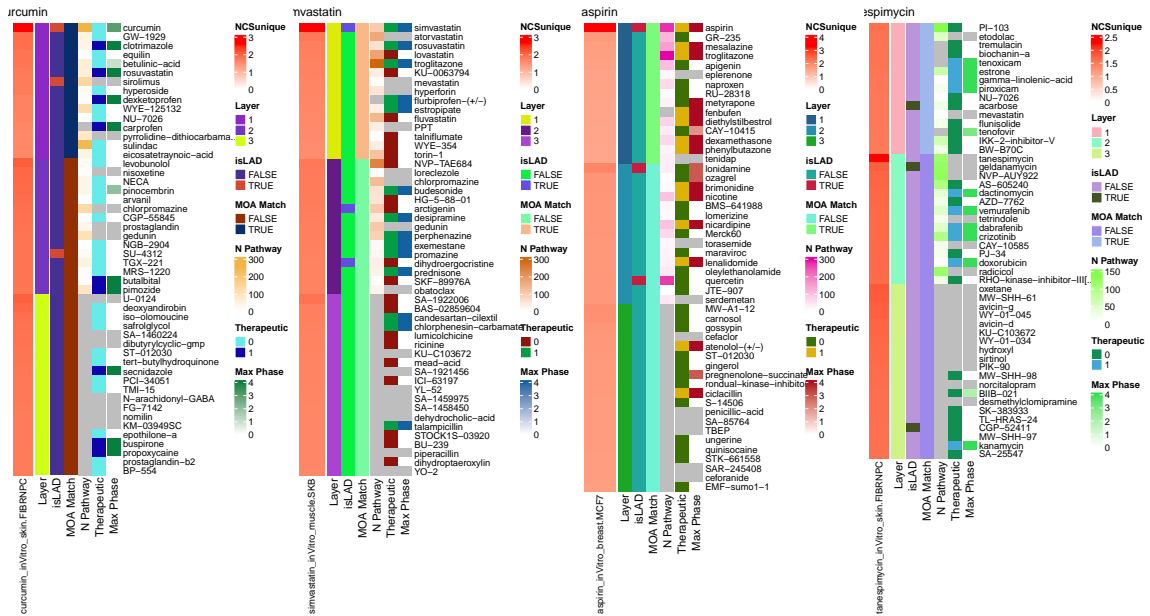


Figure S5: Top 50 PDs from voteNCSUnique method with stratification on 4 query LADs of curcumin, simvastatin, aspirin, tansespimycin individually. The legends are the same as Figure 3.4E and 3.4F.

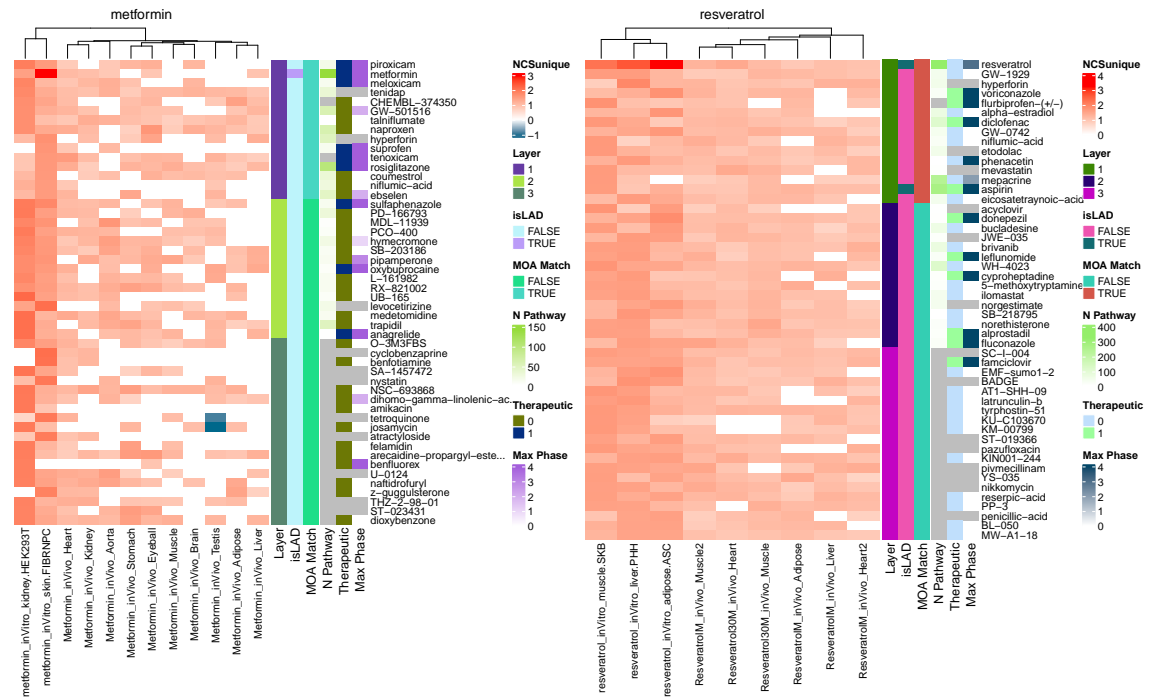


Figure S6: Top 50 PDs from voteNCSUnique method with stratification on 2 query LADs of metformin and resveratrol individually. The legends are the same as Figure 3.4E and 3.4F.

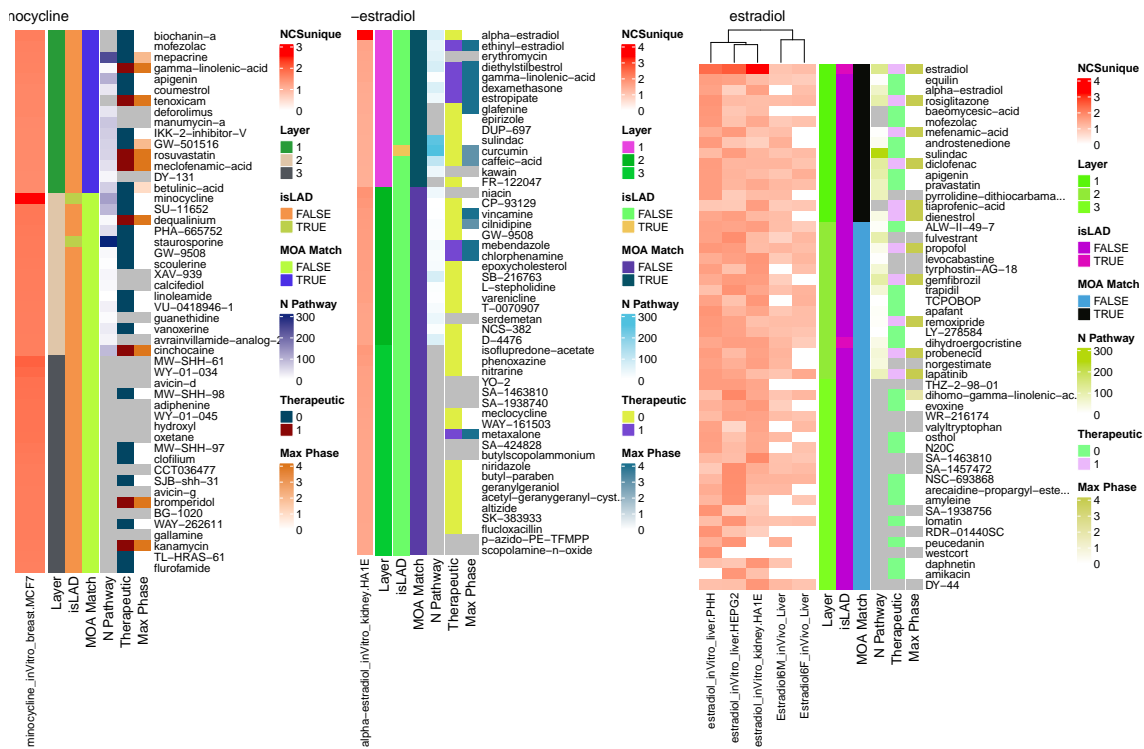


Figure S7: Top 50 PDs from voteNCSUnique method with stratification on 3 query LADs of minocycline, alpha-estradiol, estradiol individually. The legends are the same as Figure 3.4E and 3.4F.

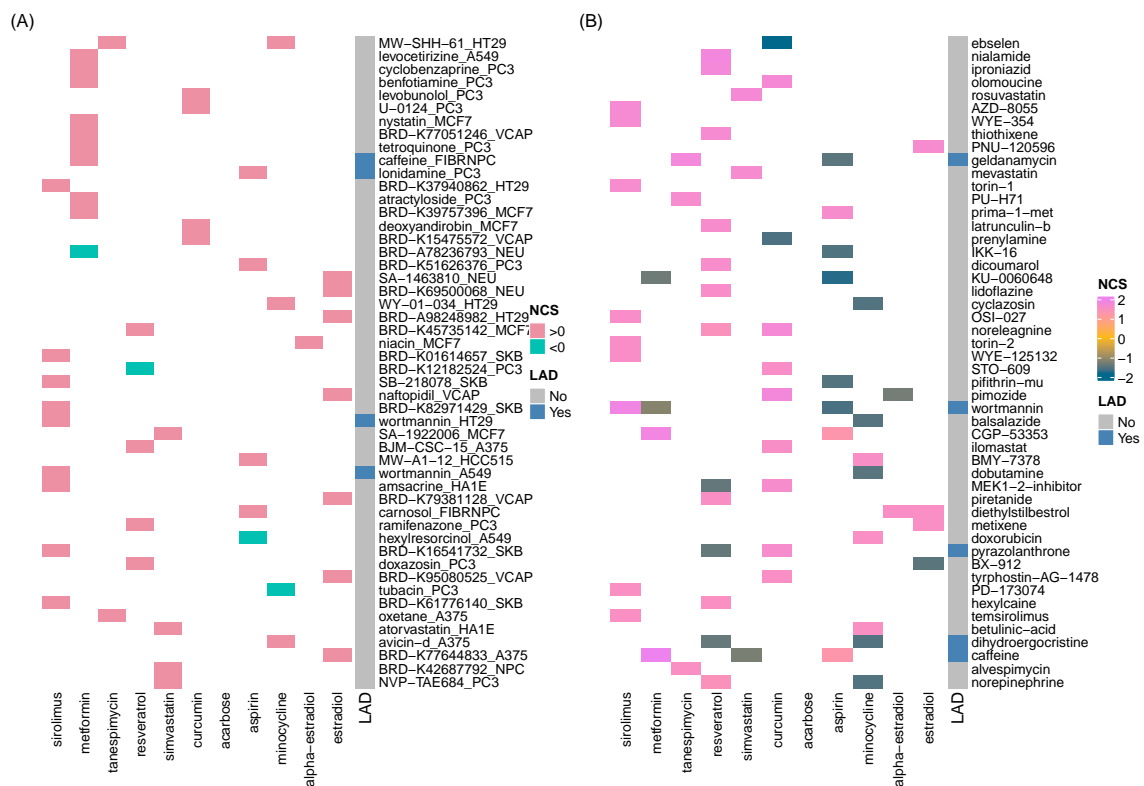


Figure S8: Heatmap summary of top 50 ranking drugs from LAD11 in vitro queries in selected cells from RRS cutoffs or target match filtering. (A) Heatmap of top 50 ranking drugs by filtering the GESS results with a minimum RRS5 of 0.6 and NCS score not equal to zero. (B) Heatmap of top 50 ranking drugs by filtering the GESS results with an NCS score not equal to 0 and selecting drugs that share at least one target with the query LAD in its GESS result. Known LADs are annotated in blue in the binary color bar to the right of the heatmap. The MOA and targets annotations of the identified known LADs are shown in Table S9.

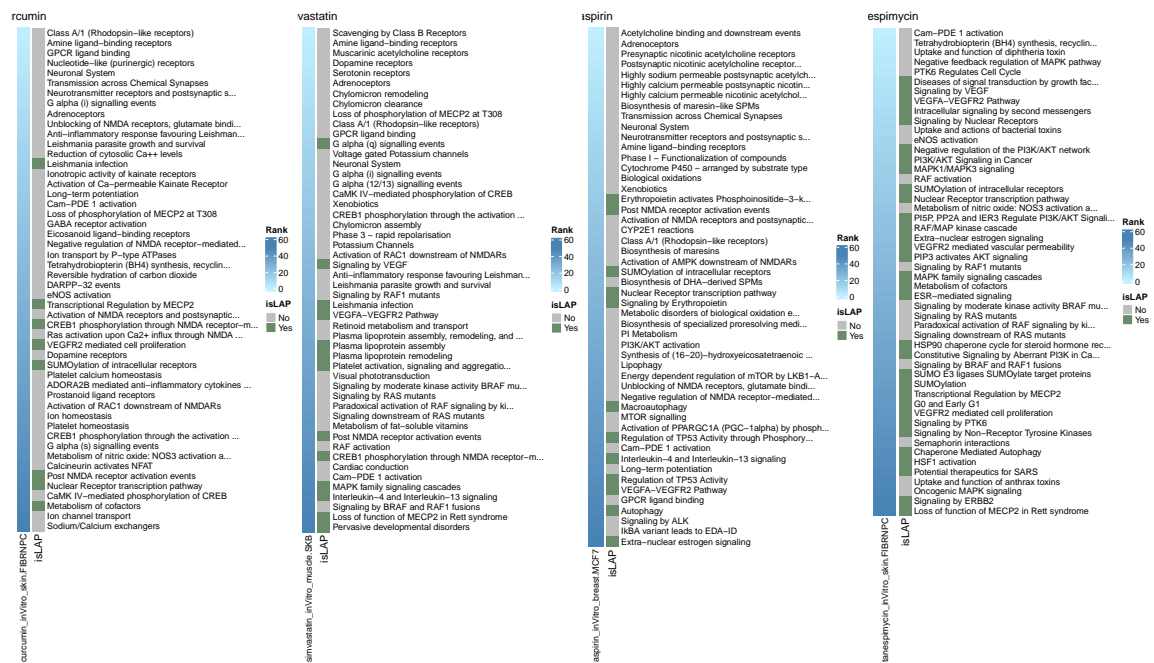


Figure S9: Top 50 PPs from vote strategy on 4 query LADs of curcumin, simvastatin, aspirin, tanspimycin individually. The legends are the same as Figure 3.6B and Figure 3.6C.

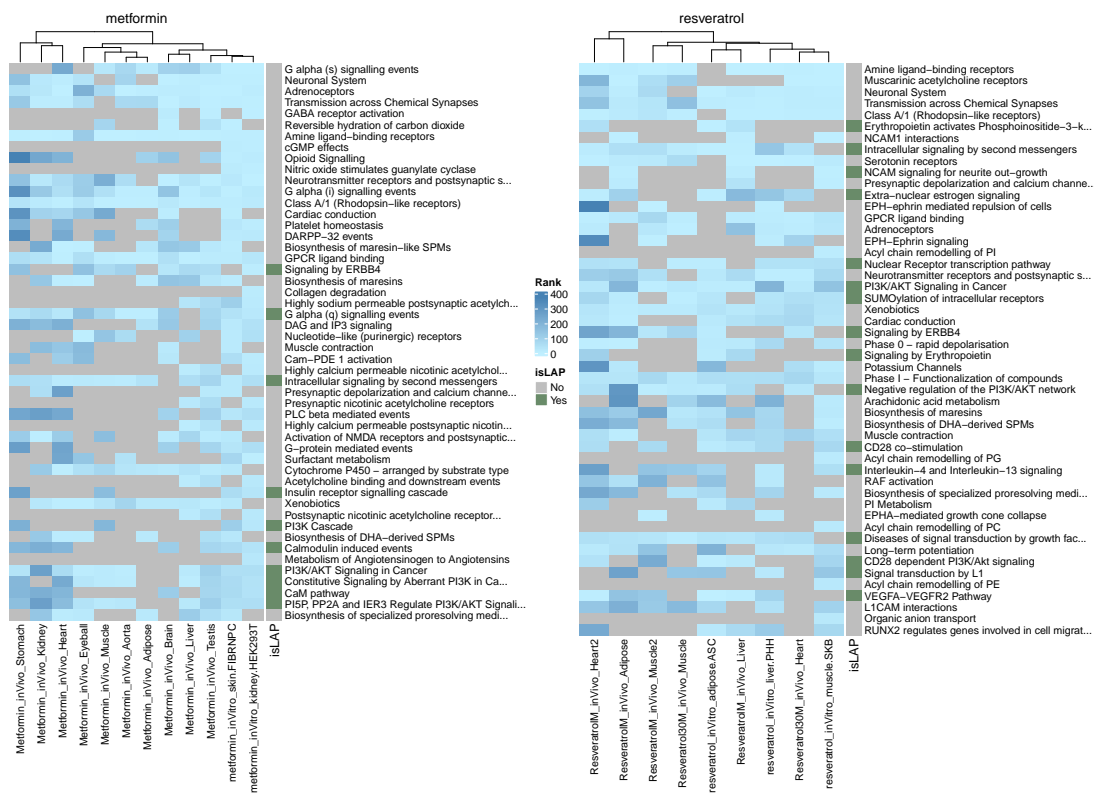


Figure S10: Top 50 PPs from vote strategy on 2 query LADs of metformin and resveratrol individually. The legends are the same as Figure 3.6B and Figure 3.6C.

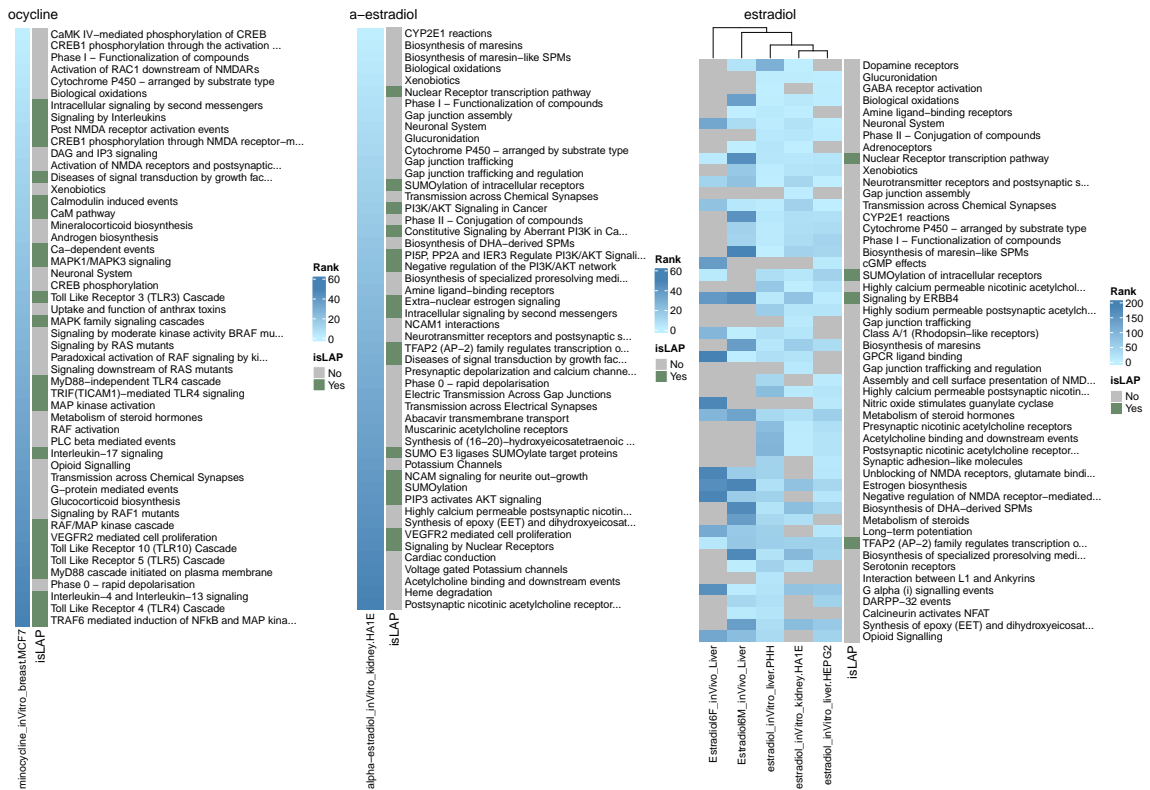


Figure S11: Top 50 PPs from vote strategy on 3 query LADs of minocycline, alpha-estradiol, estradiol individually. The legends are the same as Figure 3.6B and Figure 3.6C.

Chapter 4

Annotation Packages

4.1 Abstract

To enhance the GESS results generated by *signatureSearch* with detailed functional annotation data of both drugs and targets, I developed several affiliated data packages. The required annotations were collected from DrugBank, DrugAge, CMAP2 and LINCS. As user-friendly query interface I developed for this purpose another R/Bioconductor package called *customCMPdb*. An important feature of this package is support for custom compound collections. It supports querying the annotations across the pre-built or custom annotation database by providing any type of query compound IDs. I also developed the *drugbankR* package for drug-target interaction annotations from DrugBank database. It also provides utility to show whether the drugs are FDA approved. A Shiny web application named as *geneTargetAnno* was also built to support getting the drug-target annotations from community databases including DrugBank and STITCH online. This web service allows to get targeted drugs with structures for the query genes with Ensembl gene IDs, which also in-

cludes the target protein/gene IDs of the result drugs. The uniform resource locator (URL) links to this web application with query gene IDs can be easily added to users' existing gene tables as an extension of their drug-target annotations. For example, the gene table could be the differential expressed genes that are affected by the biological state of interest. In addition to the *signatureSearch* package, it serve as a complementary way to identify candidate drugs and potential treatments for diseases.

4.2 Materials and Methods

4.2.1 Implementation

The R/Bioconductor packages such as *drugbankR* and *customCMPdb* have been implemented as an open-source Bioconductor package using the R programming language for drug-target interaction, compounds and structures annotations. The pre-built compound annotation and structure databases are hosted on Bioconductor's AnnotationHub. Both packages are freely available for all common operating systems. To optimize reusability and performance, their functions and data containers are designed based on existing Bioconductor S4 core classes. The up-to-date source locations and versions of data sets are provided in the vignettes and help files of the two packages.

The Shiny web application was built straightly from R package as the backend of the application. The Shiny app was first created by developing the User Interface (UI), which is achieved by adapting the structure of the code to the requirements of the Shiny app structure. The Shiny environment was used to make the frontend communicate with the functions and objects in the backend environments. This can be done with the `shinyApp()`

function which takes the object defining the UI interface and the function that takes the input and delivers the output. The Shiny application can be easily launched locally by executing the `runApp()` code or clicking the ‘Run App’ button at the top of users Shiny app code in the RStudio environment. The local application was then deployed and hosted on the shinyapps.io server as an online web service.

4.2.2 Software Design

Integrating compound annotations including drug-target interactions and structural information in DrugBank, STITCH, DrugAge, CMAP2 and LINCS databases in a single environment as R/Bioconductor packages has several advantages. First, Bioconductor provides access to a large number of compound analysis tools that are interoperable by sharing the same data structures and S4 classes optimized for compound analysis. Second, it consolidates an expandable number of community compound annotations and structure information into a single environment along with options to add customized compound annotations. It also allows users to easily query the compound annotations across any selected databases by providing any type of the query compound IDs. Third, the usage of generic data objects and classes improves maintainability and reproducibility of the provided functionalities, while the integration with the existing R/Bioconductor ecosystem, such as the widely used `SDFset` S4 object in *ChemmineR* package, maximizes their extensibility and reusability for other compound analysis applications.

4.3 Results

4.3.1 drugbankR Package

The *drugbankR* package was created to get the drug annotations (mainly drug-target annotation) in DrugBank database through R. The source code is shared at [GitHub](#), the package can be directly installed from GitHub and loaded into R session by running

```
devtools::install_github("yduan004/drugbankR")
library(drugbankR)
```

This package can be used to query any downloadable version of the DrugBank database in R. The latest version at the time of writing this thesis is 5.1.8. The downloaded and unzipped DrugBank database in XML file format can be loaded into R session as a data.frame object by passing the `xmlfile` argument of the `dbxml2df` function the path to the downloaded XML file, and running code in R session as

```
dbdf <- dbxml2df(xmlfile="drugbank_5.1.8.xml", version="5.1.8")
```

This process may take about 20 minutes. The `version` argument is a character indicating the version of the downloaded DrugBank database. Users need to create a free DrugBank account and log in to download the data. The generated `dbdf` data.frame object can be stored into an SQLite database by running

```
df2SQLite(dbdf=dbdf, version="5.1.8")
```

The generated DrugBank SQLite database in the specified version is stored under user's current working directory of R session named as `drugbank_5.1.8.db`. Users can check their current R working directory by running `getwd()` in R. The SQLite database can be queried by the `queryDB` function. One can get the entire DrugBank data frame by running

```
|| dbdf <- queryDB(type="getAll", db_path="drugbank_5.1.8.db")
```

The returned `dbdf` object is a `data.frame` with 14,315 rows and 55 columns. Each row represents a drug entry in DrugBank with DrugBank ID. The columns are drug annotations such as name, description, CAS number, state, groups, indication, pharmacodynamics, mechanism of action, toxicity, half life, protein binding, classification, international brands, manufacturers, prices, dosages, FDA label, external identifiers, pathways, targets. Table 4.1 shows the six drug entries with several selected annotations as an example. One can also retrieve all the valid DrugBank ids by running

```
|| ids <- queryDB(type="getIDs", db_path="drugbank_5.1.8.db")
```

The output `ids` variable is a character vector of DrugBank IDs with drug names in the `names` slot. There are a total of 14,315 valid DrugBank IDs/entries in the 5.1.8 version. One can also determine whether the drugs are FDA approved by inputting the query DrugBank IDs and running

```
|| queryDB(ids=c("DB00001", "DB00002", "DB00111"), type="whichFDA", db_
  path="drugbank_5.1.8.db")
```

The three drugs passed to the `ids` argument are used as an example. The output is a `data.frame` object indicating the name of the query IDs and whether they are FDA approved with logic values as shown in Table 4.2. Finally, one can get the gene/protein target ID systems including DrugBank id, UniProt id, UniProt name and gene symbol of the query drugs by running

```
|| queryDB(ids=c("DB00001", "DB00002", "DB00111"), type="getTargets",
  db_path="drugbank_5.1.8.db")
```


The output is shown in Table 4.3. For queries that have many targets, only three out of them are shown here and the others are deleted and replaced with three dots for better display.

Table 4.1: Six drug entries in DrugBank database with several selected column annotations as an example.

drugbank-id	name	cas-number	state	groups	manufacturers
DB00001	Lepirudin	138068-37-8	liquid	approved	Bayer healthcare pharmaceuticals inc
DB00002	Cetuximab	205923-56-4	liquid	approved	
DB00003	Dornase alfa	143831-71-4	liquid	approved	Genentech, Inc
DB00004	Denileukin diftitox	173146-27-5	liquid	approvedinvestigational	
DB00005	Etanercept	185243-69-0	liquid	approvedinvestigational	Amgen Inc. + Wyeth + Takeda
DB00006	Bivalirudin	128270-60-0	solid	approvedinvestigational	The medicines co

Table 4.2: Annotation table of the name of the query DrugBank IDs and whether they are FDA approved with logic values.

DrugBank ID	name	whichFDA
DB00001	Lepirudin	TRUE
DB00002	Cetuximab	TRUE
DB00111	Daclizumab	FALSE

4.3.2 Shiny Web Application for Gene Target Annotation

To get drug-target annotations for query genes conveniently online, a Shiny web application named *geneTargetAnno* was developed. This application can be used to get the targeted drugs for the query Ensembl gene IDs. It also shows the structures and target protein/gene IDs of the result drugs. The URL links to this web application with specified query gene IDs can be easily added to users' existing gene tables as an extension of their drug-target annotations, such as the gene list that are affected by human longevity associated SNPs. The gene target annotation web service serves as a complementary way to identify candidate drugs for genes associated with a biological state of interest. The source code

Table 4.3: Target annotation table of the query drugs. The queries are DrugBank IDs (Q-DBID). Their gene/protein targets have four ID systems including DrugBank target ID (T-DBID), UniProt ID (T-UnipID), UniProt name (T-UnipName) and gene symbol (T-Gene). For queries that have many targets, only three out of them are shown. The others are deleted and replaced with three dots for better display.

Q-DBID	T-DBID	T-UnipID	T-UnipName	T-Gene
DB00001	BE0000048	P00734	Prothrombin	F2
DB00002	BE0000767;	P00533;	Epidermal growth factor receptor; Low affin-	EGFR;
	BE0000901;	O75015;	ity immunoglobulin gamma Fc region receptor	FCGR3B;
	BE0002094	P00736 ...	III-B; Complement C1r subcomponent ...	C1R ...
	...			
DB00111	BE0000658;	P01589;	Interleukin-2 receptor subunit alpha;	IL2RA;
	BE0000651;	P14784;	Interleukin-2 receptor subunit beta; Low	IL2RB;
	BE0000901	O75015	affinity immunoglobulin gamma Fc region	FCGR3B
	receptor III-B

of the *geneTargetAnno* application is shared at [GitHub](#). It can be deployed locally by the `runApp` function from the *shiny* R/Bioconductor package.

Figure 4.1 is the screenshot of the *geneTargetAnno* web interface that shows the gene target annotation results in DrugBank database for the query Ensembl gene ID. The main utilities of this web application are marked in red symbols. The input Ensembl gene ID can be in the URL address by specifying the value of `database` and `symbol` in the format of `?database=DrugBank&symbol=ENSG00000124275` that is appended to `https://tgirke.shinyapps.io/geneTargetAnno/` (input 1) and then refreshing the page. It can also be typed in the submission text box (input 2) and users can hit the **Submit** button to submit the gene id. The gene-target annotation result tables in DrugBank and STITCH databases will be shown under the two navigation tabs (Figure 4.2). Users can switch between them by clicking the tab. The DrugBank annotation table include the UniProt ID of the query gene. It also includes the drugs that target the query gene, as well as the structures and targets of the drugs. The STITCH annotation table (Figure 4.2)

The screenshot shows the web interface for gene target annotation. The URL in the browser is `tgirke.shinyapps.io/geneTargetAnno/?database=DrugBank&symbol=ENSG00000124275`. The page title is "Drugs and Target Proteins". On the left, there is a form with "Ensembl Gene ID" and a text input field containing "ENSG00000124275", with a "Submit" button below it. Above the form, there are two database selection buttons: "DrugBank" and "STITCH". The main heading is "Drugs and target proteins for ENSG00000124275". Below this, there is a "Show" dropdown menu set to "entries". A table header has four columns: "Uniprot_id", "target_drugs", "structure", and "drug_targets". Each column has a dropdown menu set to "All". The table contains two rows of results. Row 1 shows UniProt ID "Q9UBK8", target drug "DB00115", a complex drug structure, and drug targets "Q99707 [Homo sapiens (Human)]", "P22033 [Homo sapiens (Human)]", "Q9Y4U1 [Homo sapiens (Human)]", and "P42898 [Homo sapiens (Human)]". Row 2 shows UniProt ID "Q9UBK8", target drug "DB00134", a chemical structure of a thiolamide, and drug targets "Q9UBK8 [Homo sapiens (Human)]", "Q99707 [Homo sapiens (Human)]", and "Q9H2M3 [Homo sapiens (Human)]".

Figure 4.1: Screenshot of the *geneTargetAnno* web interface that shows the gene target annotation results in DrugBank database for the query Ensembl gene ID. The main utilities were marked in red symbols. The check marks (number 4) show the columns in the DrugBank annotation table including UniProt ID of the query gene (*Uniprot_id*), DrugBank IDs of the target drugs (*target_drugs*), drug structures (*structure*) and UniProt IDs of the drug targets (*drug_targets*). The brackets in the *drug_targets* column also include the supported species of the drug-target interaction.

include the Ensembl protein IDs of the input Ensembl gene ID, PubChem CIDs of the target drugs, drug structures, and Ensembl protein IDs of drug targets. Every genes, drugs and proteins in the website are clickable and can be linked to the corresponding entries from the official website. For example, the Ensembl gene IDs can be linked to the gene card page in the official Ensembl website at <http://www.ensembl.org/>. The full gene-target annotation result table for the DrugBank database is shown in Figure 4.3. It can be searched at the search box. Each column can be ranked alphabetically or searched. The full table can be downloaded as csv, Excel or pdf file.

← → ↻ tgirke.shinyapps.io/geneTargetAnno/?database=DrugBank&symbol=ENSG00000124275 ☆

Drugs and Target Proteins

Ensembl Gene ID

DrugBank STITCH

Drugs and target proteins for ENSG00000124275

Show entries

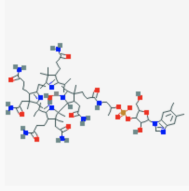
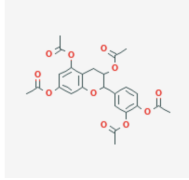
	ensembl_protein_id	target_drugs	structure	drug_targets
	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>
1	ENSP00000264668	04479097		ENSP00000215838, ENSP00000257264, ENSP00000264668, ENSP00000355536, ENSP00000367064, ENSP0000038384
2	ENSP00000264668	03363314		ENSP00000264668, ENSP00000297494, ENSP0000032725

Figure 4.2: Screenshot of the *geneTargetAnno* web interface that shows the gene target annotation results in the STITCH database. The columns in the STITCH annotation table include the Ensembl protein IDs of the input Ensembl gene ID (*ensembl_protein_id*), PubChem CIDs of the target drugs (*target_drugs*), drug structures (*structure*) and Ensembl protein IDs of drug targets (*drug_targets*)

4.3.3 customCMPdb package

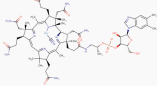
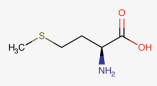
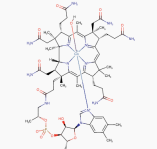
I also built an R/Bioconductor package named *customCMPdb* to integrate the community and custom compound collections. This package serves as a query interface for important community collections of small molecules, while also allowing users to include custom compound collections. Both annotation and structure information is provided. The annotation data is stored in an SQLite database, while the structure information is stored in Structure Definition Files (SDF). Both are hosted on Bioconductor's *AnnotationHub*. This package has been already published and is available on Bioconductor at [here](#).

Drugs and target proteins for **ENSG00000124275**

Show entries 1. Search:

Uniprot_id ² target_drugs structure drug_targets

3.

1	Q9UBK8	DB00115		Q99707 [Homo sapiens (Human)] P22033 [Homo sapiens (Human)] Q9UBK8 [Homo sapiens (Human)] Q8IVH4 [Homo sapiens (Human)] Q9Y4U1 [Homo sapiens (Human)] P42898 [Homo sapiens (Human)]
2	Q9UBK8	DB00134		Q9UBK8 [Homo sapiens (Human)] Q99707 [Homo sapiens (Human)] P50579 [Homo sapiens (Human)] Q93088 [Homo sapiens (Human)] Q9H2M3 [Homo sapiens (Human)]
3	Q9UBK8	DB00200		Q99707 [Homo sapiens (Human)] P22033 [Homo sapiens (Human)] Q9UBK8 [Homo sapiens (Human)] Q8IVH4 [Homo sapiens (Human)] P20061 [Homo sapiens (Human)] Q9BXJ7 [Homo sapiens (Human)] O60494 [Homo sapiens (Human)] Q96EY8 [Homo sapiens (Human)] Q9Y4U1 [Homo sapiens (Human)]

4. Showing 1 to 3 of 3 entries Previous Next

Figure 4.3: Full gene-target annotation result table for the DrugBank database. It can be searched at the search box (number 1). Each column can be ranked alphabetically (number 2) or searched (number 3). The full table can be downloaded as csv, Excel or pdf file via the buttons at number 4.

Pre-configured Databases

The following is the description of the four annotation tables stored in the pre-built SQLite Database. The DrugAge database was downloaded from [here](#) as a CSV file. The downloaded `drugage.csv` file contains annotation columns such as compound name, species, strain, dosage, average lifespan change, gender, significance, and pubmed id. Since the DrugAge database only contains the drug name as identifiers, it is necessary to map the drug name to other uniform drug identifiers, such as ChEMBL IDs. In the *custom-CMPdb* package, the drug names have been mapped to [ChEMBL](#) [61], [PubChem](#) [101] and DrugBank IDs semi-manually and stored under the `inst/extdata` directory named as

`drugage_id_mapping.tsv`. Part of the id mappings in the `drugage_id_mapping.tsv` table was generated by the `processDrugage` function by strict matching of the compound name in DrugAge to the `pref_name` in the ChEMBL database (version 24). The missing IDs were added manually. A semi-manual approach was to use this [web service](#) by entering the compound name as Chemical Name and choosing to convert to ChEMBL IDs. After the semi-manual process, the left ones were manually mapped to ChEMBL, PubChem and DrugBank ids. The mixture entries such as green tee extract or peptide (*e.g.* Bacitracin) were commented. The `drugage_id_mapping` table was then built into the annotation SQLite database named as `compoundCollection_0.1.db` by the `buildDrugAgeDB` function.

The DrugBank annotation table was transformed from the downloaded DrugBank database in [xml file](#). The extracted xml file was processed by the `dbxml2df` function in the *drugbankR* package. The `dbxml2df` and `df2SQLite` functions in the package were used to load the xml file into R as a `data.frame` R object and then store the `data.frame` in the SQLite database named as `compoundCollection_0.1.db`. The DrugBank table has the comprehensive annotation columns, such as drugbank id, name, description, CAS number, indication, pharmacodynamics, mechanism of action, toxicity, metabolism, half life, protein binding, classification, synonyms, international brands, packagers, manufacturers, prices, dosages, FDA label, pathways, and targets. The DrugBank ID to ChEMBL ID mappings were obtained from [UniChem](#).

The CMAP2 annotation table was downloaded from the [instance table](#) at Broad Institute and processed by the `buildCMAPdb` function in the *customCMPdb* package. The CMAP2 instance table contains the drug annotation columns, such as instance id, batch

id, cmap name, concentration (Molar), duration (hour), cell, array, perturbation scan id, vehicle scan id, scanner, vehicle, vendor, and catalog number. The `buildCMAPdb` function maps the drug names to external drug IDs including UniProt [192], PubChem, DrugBank and ChemBank [166]. It also adds additional annotation columns such as directionality, ATC codes, and SMILES string. The generated `cmap.db` SQLite database from the `buildCMAPdb` function contains both the compound annotation table and structure information. The ChEMBL id mappings were further added to the annotation table via the PubChem CID to ChEMBL id mappings from [UniChem](#). The CMAP2 annotation table was stored in the `compoundCollection` SQLite annotation database.

The LINCS compound annotation table was downloaded from [GEO](#) where only compound treatment type was selected. The annotation columns include lincs id, perturbation name, is touchstone, inchi key, canonical SMILES, and pubchem cid. The annotation table was stored in the `compoundCollection` SQLite database. Since the annotation only contains LINCS id to PubChem CID mapping, the LINCS IDs were also mapped to ChEMBL IDs via inchi key. At the time of writing, the following community databases were included in the package: [DrugAge](#) [6], [DrugBank](#) [215], [CMAP2](#) [109], and [LINCS](#) [186].

In addition to providing access to the above pre-built collection of compound annotations and structures, the package also supports the integration of custom collections of compounds, which will be automatically stored for the user in the same data structure as the pre-configured databases. Both custom collections and those provided by this package can be queried in a uniform manner, and then further analyzed with cheminformatics packages such as *ChemmineR*, where SDF files are imported into R as flexible S4 containers [23].

The compounds annotation tables for the pre-configured four databases (DrugAge, DrugBank, CMAP2 and LINCS) are stored in an cached SQLite database via AnnotationHub. The cached SQLite database can be loaded into a user's R session by running code of

```
|| conn <- loadAnnot()
```

The annotation tables can be queried by running `dbReadTable` function on the SQLite connection and selecting the corresponding table name. For example the DrugAge annotation table is named as `drugAgeAnnot`. Table 4.4 shows the top six rows of the DrugAge annotation table with several selected columns as an example.

The compound structures in the above four databases can be obtained by loading their corresponding SDF files into R as an `SDFset` object. Each database has its corresponding SDF files stored in AnnotationHub where the files can be downloaded in the cached folder of user's local computer. The path to the cached DrugAge SDF file can be obtained via the AnnotationHub ID of AH79564. The `read.SDFset` function from the *ChemmineR* package can be used to read the SDF file into R as an `SDFset` object. Utilities from the *ChemmineR* package including the plotting can be used on the loaded DrugAge `SDFset` object. Instructions on how to work with `SDFset` objects are provided in the ChemmineR [vignette](#). For instance, one can plot any of the loaded structures with the `plot` function. The structures of three out of six compounds in Table 4.4 that have the corresponding SDF instances are plotted in Figure 4.4.

The SDF file for drugs in the DrugBank database can be loaded into R in the same way. The corresponding AnnotationHub ID of the DrugBank SDF file is AH79565.

Table 4.4: Top six rows of the DrugAge annotation table with several selected columns. Avg: average lifespan change, PCID: PubChem CID, DBID: DrugBank ID.

DrugAge ID	Drug Name	Species	Strain	Dosage	Avg	Gender	PCID	DBID
ida00001	Vitexin	Caenorhabditis elegans	N2	50 μ M	8		5280441	
ida00002	Cyclosporin A	Caenorhabditis elegans		88 μ M	18			DB00091
ida00003	Histidine	Caenorhabditis elegans	N2	5 mM	10		6274, 6971009	DB00117
ida00004	SRT1720	Mus musculus	C57BL/6J	100 mg/kg body weight	8.8			
ida00005	Cordyceps sinensis oral liquid	Drosophila melanogaster	Oregon-K	0.20 mg/ml	32	Male		
ida00006	Lysine	Caenorhabditis elegans	N2	5 mM	8		5962, 122198194	DB00123, DB11101

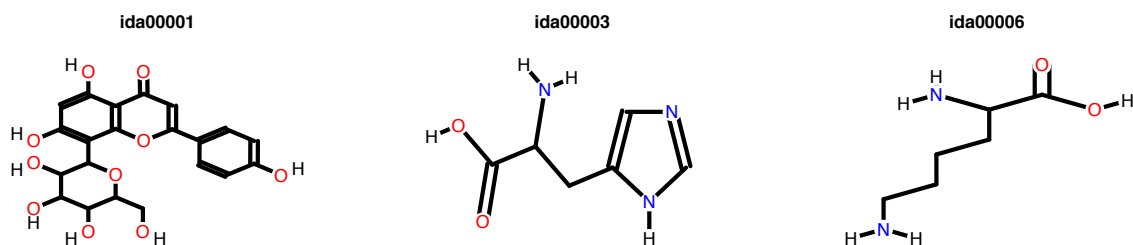


Figure 4.4: Structures of the three out of six compounds in Table 4.4 that have the corresponding SDF instances.

The SDF file was obtained by downloading from the [DrugBank](#) website. During the import into R, *ChemmineR* checks the validity of the imported compounds. The AnnotationHub IDs of the cached CMAP2 and LINCS SDF files are AH79566 and AH79567, respectively. The import of the SDF files works the same way. For reproducibility, the R code for generating the SQLite annotation database and the above four SDF files is included in the `inst/scripts/make-data.R` file of the *customCMPdb* package. The file location on a user's system can be obtained with R code of

```
||| system.file("scripts/make-data.R", package="customCMPdb")
```

Custom Annotation Database

As previously mentioned, the pre-configured SQLite annotation database is hosted on Bioconductor's *AnnotationHub*. Users can download it to a local *AnnotationHub* cache directory. The path to the cached database can be obtained by running

```
ah <- AnnotationHub()
annot_path <- ah[["AH79563"]]
```

Users can also add their custom compound annotation tables to the cached SQLite database via the `addCustomAnnot` function in this package. The following introduces how users can import to the SQLite database their own compound annotation tables. In this case, the custom annotation data needs to be a `data.frame` related object and the ChEMBL IDs need to be included under the column named `chembl_id`. The name of the custom data table is case-insensitive and can be specified under the `annot_name` argument of the `addCustomAnnot` function. The custom annotation tables can also be deleted by referencing their names via the `deleteAnnot` function. The `listAnnot` function can be used to obtain a list of the existing annotation tables in the SQLite database. The SQLite database can also be set back to the pre-built version via the `defaultAnnot` function. This is achieved by deleting the existing SQLite database and re-downloading a fresh instance from *AnnotationHub*.

The `queryAnnotDB` function can be used to query the compound annotations from the default resources as well as the custom resources stored in the SQLite database. The query input can be a set of ChEMBL IDs. In this case it returns a `data.frame` object containing the annotations of the matching compounds from the selected annotation resources specified under the `annot` argument. The `listAnnot` function returns the names that can

be assigned to the `annot` argument. Table 4.5 shows the annotation table of the five query compounds with ChEMBL IDs in LINCS and custom databases. Since the supported compound databases use different identifiers, a ChEMBL ID mapping table is used to connect identical entries across databases as well as to link out to other resources such as ChEMBL itself or PubChem. For custom compounds, where ChEMBL IDs are not available yet, users can use alternative and/or custom identifiers. For example, users can use LINCS IDs to query the LINCS annotation data.

Table 4.5: Annotation table of the five query compounds with ChEMBL IDs in LINCS and custom databases. `isTS`: whether the compounds are in Touchstone database, `PCID`: PubChem CID, `feature1`, `2`: two columns in the custom annotation table added by user.

ChEMBL ID	LINCS ID	Drug Name	isTS	PCID	feature1	feature2
CHEMBL10	BRD-A37704979	SB-203580	0	176155	f3	-1.88
CHEMBL1004	BRD-A44008656	doxylamine	1	-666	f2	1.56
CHEMBL1064	BRD-K22134346	simvastatin	0	-666		
CHEMBL113	BRD-K02404261	caffeine	1	-666		
CHEMBL31574					f5	0.01

4.4 Discussion

The compound annotations from different resources are usually independent from each other. Integrating or collecting drug or small molecule annotations from different sources into a single R/Bioconductor package has not been done before and it provides several unique advantages. First, the *customCMPdb* packages I developed serves as a query interface and makes it more convenient and accessible to get compound annotations from importance community platforms in a single environment. It supports getting the drug annotations across the pre-built or custom annotation databases by providing any type of the query IDs. The annotation tables from different resources were consolidated into a single

well-designed SQLite database with a master compound ID mapping table, which supports adding an extendable number of other annotation tables with reasonable sizes. Users can get compound annotations of their query compounds in both the pre-built databases and the added customized ones. Moreover, Both annotation and structure information are provided in this package and hosted on Bioconductor's *AnnotationHub*. Finally, the usage of generic data objects and classes improves maintainability and reproducibility of the provided functionalities, while the integration with the existing R/Bioconductor ecosystem maximizes their extensibility and reusability for other cheminformatic tools.

Two assistant tools were also developed mainly for drug-target interaction annotations in DrugBank and STITCH databases. The *drugbankR* package was developed specifically for compound annotations in DrugBank database. It fills the blank that currently no Application Programming Interface (API) available for drug-target annotations on the DrugBank website, no local tools (*e.g.* R packages) to access drug annotations locally in batch. It also provides utility to indicate whether the drugs are FDA approved from the DrugBank annotation. In addition to local tools, I also developed a Shiny web service named as *geneTargetAnno* to get drug-gene interaction annotations for both the DrugBank and STITCH databases online. This Shiny web application accepts an Ensembl gene ID as input either at the URL or in the input box to get its targeted drugs with structures. The query results also contain the target protein/gene IDs of the result drugs. The URL address containing the query gene IDs can be easily added to users' existing gene tables as an extension of drug-target annotations. It serves as a complementary way to identify candidate drugs and potential treatments for diseases.

In total, the three cheminformatic tools I developed contribute to the compound annotation field of cheminformatics. More annotation sources can be added if they have reasonable size that can be stored locally, such as the ChEMBL database. The latter can be accessed via the web-interface. It also supports data downloads with many database formats including Oracle, MySQL, PostgreSQL, and SQLite to query it locally.

Chapter 5

Conclusion

This thesis is divided into three main components: (1) development of efficient GESS and FEA methods, and their implementation in reusable community software; (2) application of the resulting GESS and FEA workflow to the discovery of novel healthy aging drugs; and (3) development of helper software and data packages that extend the functionalities of the GESS software.

The *signatureSearch* R/Bioconductor package provides an integrated environment for identifying similar GESs in reference databases and guiding the downstream functional interpretation of the discovered connections. The package is unique in that it includes several novel search and enrichment methods in a single environment with efficient data structures and access to pre-built GES databases. It also allows users to work with custom databases. Subsequently, I tested the performance of different GESS methods. These are the first systematic performance tests of GES search methods reported so far. The *signatureSearch* software paves the way for discovering biologically relevant connections and gain insights

into the applied biological researches, such as studies on improving treatments for diseases or identifying novel target site candidates for drugs.

The *signatureSearch* environment was then applied to the human longevity and healthy aging research filed to discover LADs, LAGs and LAPs by searching the LINCS database. I identified a list of known drugs that are currently mainly used in treating cancers and diseases. They can be used for drug repurposing as healthy aging drugs in the human longevity field. I also identified a list of small molecules under experimental studies that can induce longevity GESs. Some of them have been tested in model organisms. A list of proteins and pathways that are related to longevity and can be targeted by pharmaceutical drugs for lifespan extension strategies were also identified.

To facilitate the *signatureSearch* package to annotate the compounds from different sources, I developed several data packages that incorporate detailed annotations and structures of drugs from different community databases. The *customCMPdb* is the first package that integrates the compound collections from different communities into a single well-designed SQLite database. It also has a user-friendly query interface and supports adding custom compound collections. The *drugbankR* package is built to obtain the drug annotations from the DrugBank database locally. It addresses the limitation that currently the DrugBank annotations can only be queried online with one drug at a time. The local tool supports getting the drug annotations in batch. To obtain drug-target annotations from a web interface for any gene or protein list provided by users, I developed the Shiny web application *geneTargetAnno*. The web application URLs with query gene IDs can be easily added to users' existing gene tables as an extension of drug-target annotations. In

addition to the *signatureSearch* package, these several affiliated data packages serve as a complementary way to identify candidate drugs and potential treatments for diseases.

Bibliography

- [1] Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J T Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.*, 20(1):194, September 2019.
- [2] Oge Arum and Thomas E Johnson. Reduced expression of the caenorhabditis elegans p53 ortholog cep-1 results in increased longevity. *J. Gerontol. A Biol. Sci. Med. Sci.*, 62(9):951–959, September 2007.
- [3] Gil Atzmon, Miook Cho, Richard M Cawthon, Temuri Budagov, Micol Katz, Xiaoman Yang, Glenn Siegel, Aviv Bergman, Derek M Huffman, Clyde B Schechter, Woodring E Wright, Jerry W Shay, Nir Barzilai, Diddahally R Govindaraju, and Yousin Suh. Evolution in health and medicine sackler colloquium: Genetic variation in human telomerase is associated with telomere length in ashkenazi centenarians. *Proc. Natl. Acad. Sci. U. S. A.*, 107 Suppl 1:1710–1717, January 2010.
- [4] Tyler William H Backman and Thomas Girke. bioassayr: Cross-Target analysis of small molecule bioactivity. *J. Chem. Inf. Model.*, 56(7):1237–1242, July 2016.
- [5] Sarah E Baker, Jacqueline K Limberg, Gabrielle A Dillon, Timothy B Curry, Michael J Joyner, and Wayne T Nicholson. Aging alters the relative contributions of the sympathetic and parasympathetic nervous system to blood pressure control in women. *Hypertension*, 72(5):1236–1242, November 2018.
- [6] Diogo Barardo, Daniel Thornton, Harikrishnan Thoppil, Michael Walsh, Samim Sharifi, Susana Ferreira, Andreja Anžič, Maria Fernandes, Patrick Monteiro, Tjaša Grum, and Others. The DrugAge database of aging-related drugs. *Aging Cell*, 16(3):594–597, 2017.
- [7] Andrzej Bartke. Single-gene mutations and healthy ageing in mammals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 366(1561):28–34, January 2011.
- [8] Nir Barzilai, Jill P Crandall, Stephen B Kritchevsky, and Mark A Espeland. Metformin as a tool to target aging. *Cell Metab.*, 23(6):1060–1065, June 2016.

- [9] Johannes H Bauer, Peter C Poon, Heather Glatt-Deeley, John M Abrams, and Stephen L Helfand. Neuronal expression of p53 dominant-negative proteins in adult *drosophila melanogaster* extends life span. *Curr. Biol.*, 15(22):2063–2068, November 2005.
- [10] J S Beckman and W H Koppenol. Nitric oxide, superoxide, and peroxynitrite: the good, the bad, and ugly. *Am. J. Physiol.*, 271(5 Pt 1):C1424–37, November 1996.
- [11] Don Benjamin, Marco Colombi, Christoph Moroni, and Michael N Hall. Rapamycin passes the torch: a new generation of mTOR inhibitors. *Nat. Rev. Drug Discov.*, 10(11):868–880, October 2011.
- [12] Yoav Benjamin and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 289, 1995.
- [13] B L Bennett, D T Sasaki, B W Murray, E C O’Leary, S T Sakata, W Xu, J C Leisten, A Motiwala, S Pierce, Y Satoh, S S Bhagwat, A M Manning, and D W Anderson. SP600125, an anthrapyrazolone inhibitor of jun n-terminal kinase. *Proc. Natl. Acad. Sci. U. S. A.*, 98(24):13681–13686, November 2001.
- [14] Bernd Bischl, Michel Lang, Olaf Mersmann, Jörg Rahnenführer, and Claus Weihs. BatchJobs and BatchExperiments: Abstraction mechanisms for using R in batch environments. *Journal of Statistical Software, Articles*, 64(11):1–25, 2015.
- [15] Mikhail V Blagosklonny. An anti-aging drug today: from senescence-promoting genes to anti-aging pill. *Drug Discov. Today*, 12(5-6):218–224, March 2007.
- [16] Mikhail V Blagosklonny. Rapamycin, proliferation and geroconversion to senescence. *Cell Cycle*, 17(24):2655–2665, December 2018.
- [17] Mikhail V Blagosklonny. Rapamycin for longevity: opinion article. *Aging*, 11(19):8048–8067, October 2019.
- [18] Marcel Bonay, Catherine Bancal, and Bruno Crestani. The risk/benefit of inhaled corticosteroids in chronic obstructive pulmonary disease. *Expert Opin. Drug Saf.*, 4(2):251–271, March 2005.
- [19] Dan Buettner and Sam Skemp. Blue zones: Lessons from the world’s longest lived. *Am. J. Lifestyle Med.*, 10(5):318–321, September 2016.
- [20] Kristopher Burkewitz, Yue Zhang, and William B Mair. AMPK at the nexus of energetics and aging. *Cell Metab.*, 20(1):10–25, July 2014.
- [21] Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas Pan-Cancer analysis project. *Nat. Genet.*, 45(10):1113–1120, October 2013.

- [22] Laura Cantini, Laurence Calzone, Loredana Martignetti, Mattias Rydenfelt, Nils Blüthgen, Emmanuel Barillot, and Andrei Zinovyev. Classification of gene signatures for their information value and functional redundancy. *NPJ Syst Biol Appl*, 4:2, 2018.
- [23] Yiqun Cao, Anna Charisi, Li-Chang Cheng, Tao Jiang, and Thomas Girke. ChemmineR: a compound mining framework for R. *Bioinformatics*, 24(15):1733–1734, August 2008.
- [24] Maurizio Cardelli, Luca Cavallone, Francesca Marchegiani, Fabiola Oliveri, Serena Dato, Alberto Montesanto, Francesco Lescai, Rosamaria Lisa, Giovanna De Benedictis, and Claudio Franceschi. A genetic-demographic approach reveals male-specific association between survival and tumor necrosis factor (A/G)-308 polymorphism. *J. Gerontol. A Biol. Sci. Med. Sci.*, 63(5):454–460, May 2008.
- [25] M Castro Cabezas, T W de Bruin, H W de Valk, C C Shoulders, H Jansen, and D Willem Erkelens. Impaired fatty acid metabolism in familial combined hyperlipidemia. a mechanism associating hepatic apolipoprotein B overproduction and insulin resistance. *J. Clin. Invest.*, 92(1):160–168, July 1993.
- [26] Alexandra Chadt and Hadi Al-Hasani. Glucose transporters in adipose tissue, liver, and skeletal muscle in metabolic health and disease. *Pflugers Arch.*, 472(9):1273–1298, September 2020.
- [27] F Chang, J T Lee, P M Navolanic, L S Steelman, J G Shelton, W L Blalock, R A Franklin, and J A McCubrey. Involvement of PI3K/Akt pathway in cell cycle progression, apoptosis, and neoplastic transformation: a target for cancer chemotherapy. *Leukemia*, 17(3):590–603, March 2003.
- [28] Jeffrey T Chang, Michael L Gatz, Joseph E Lucas, William T Barry, Peyton Vaughn, and Joseph R Nevins. SIGNATURE: a workbench for gene expression signature analysis. *BMC Bioinformatics*, 12:443, November 2011.
- [29] Hengyi Chen, Yubo Wang, Caiyu Lin, Conghua Lu, Rui Han, Lin Jiao, Li Li, and Yong He. Vorinostat and metformin sensitize EGFR-TKI resistant NSCLC cells via BIM-dependent apoptosis induction. *Oncotarget*, 8(55):93825–93838, November 2017.
- [30] Mary Chester Wasko, Abhijit Dasgupta, Genevieve Ilse Sears, James F Fries, and Michael M Ward. Prednisone use and risk of mortality in patients with rheumatoid arthritis: Moderation by use of Disease-Modifying antirheumatic drugs. *Arthritis Care Res.*, 68(5):706–710, May 2016.
- [31] Arnab Roy Chowdhury, Suparna Mandal, Bidyottam Mittra, Shalini Sharma, Sibabrata Mukhopadhyay, and Hemanta K Majumder. Betulinic acid, a potent inhibitor of eukaryotic topoisomerase i: identification of the inhibitory step, the major functional group responsible and development of more potent derivatives. *Med. Sci. Monit.*, 8(7):BR254–65, July 2002.

- [32] Kaare Christensen, Thomas E Johnson, and James W Vaupel. The quest for genetic determinants of human longevity: challenges and insights. *Nat. Rev. Genet.*, 7(6):436–448, June 2006.
- [33] E J Cobos, J M Entrena, F R Nieto, C M Cendán, and E Del Pozo. Pharmacology and therapeutic potential of sigma(1) receptor ligands. *Curr. Neuropharmacol.*, 6(4):344–366, December 2008.
- [34] Steven M Corsello, Joshua A Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi E Hirschman, Stephen E Johnston, Anita Vrcic, Bang Wong, Mariya Khan, Jacob Asiedu, Rajiv Narayan, Christopher C Mader, Aravind Subramanian, and Todd R Golub. The drug repurposing hub: a next-generation drug library and information resource. *Nat. Med.*, 23(4):405–408, April 2017.
- [35] Eileen M Crimmins. Lifespan and healthspan: Past, present, and promise. *Gerontologist*, 55(6):901–911, December 2015.
- [36] Aedín C Culhane, Markus S Schröder, Razvan Sultana, Shaita C Picard, Enzo N Martinelli, Caroline Kelly, Benjamin Haibe-Kains, Misha Kapushesky, Anne-Alyssa St Pierre, William Flahive, Kermshlise C Picard, Daniel Gusenleitner, Gerald Pappenhausen, Niall O’Connor, Mick Correll, and John Quackenbush. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Res.*, 40(Database issue):D1060–6, January 2012.
- [37] Dennise D Dalma-Weiszhausz, Janet Warrington, Eugene Y Tanimoto, and C Garrett Miyada. The affymetrix GeneChip platform: an overview. *Methods Enzymol.*, 410:3–28, 2006.
- [38] Bhaskar C Das, Pritam Thapa, Radha Karki, Sasmita Das, Sweta Mahapatra, Ting-Chun Liu, Ingrid Torregroza, Darren P Wallace, Suman Kambhampati, Peter Van Veldhuizen, Amit Verma, Swapan K Ray, and Todd Evans. Retinoic acid signaling pathways in development and diseases. *Bioorg. Med. Chem.*, 22(2):673–683, January 2014.
- [39] Birgit Debrabant, Mette Soerensen, Friederike Flachsbart, Serena Dato, Jonas Mengel-From, Tinna Stevnsner, Vilhelm A Bohr, Torben A Kruse, Stefan Schreiber, Almut Nebel, Kaare Christensen, Qihua Tan, and Lene Christiansen. Human longevity and variation in DNA damage response and repair: study of the contribution of sub-processes using competitive gene-set analysis. *Eur. J. Hum. Genet.*, 22(9):1131–1136, September 2014.
- [40] Danilo Di Bona, Sonya Vasto, Cristiano Capurso, Lene Christiansen, Luca Deiana, Claudio Franceschi, Mikko Hurme, Eugenio Mocchegiani, Maeve Rea, Domenico Lio, Giuseppina Candore, and Calogero Caruso. Effect of interleukin-6 polymorphisms on human longevity: a systematic review and meta-analysis. *Ageing Res. Rev.*, 8(1):36–42, January 2009.

- [41] Adele Di Matteo, Mimma Franceschini, Sara Chiarella, Serena Rocchio, Carlo Travaglini-Allocatelli, and Luca Federici. Molecules that target nucleophosmin for cancer treatment: an update. *Oncotarget*, 7(28):44821–44840, July 2016.
- [42] Sorin Drăghici, Purvesh Khatri, Rui P Martins, G Charles Ostermeier, and Stephen A Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, February 2003.
- [43] Qiaonan Duan, St Patrick Reid, Neil R Clark, Zichen Wang, Nicolas F Fernandez, Andrew D Rouillard, Ben Readhead, Sarah R Tritsch, Rachel Hodos, Marc Hafner, Mario Niepel, Peter K Sorger, Joel T Dudley, Sina Bavari, Rekha G Panchal, and Avi Ma’ayan. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *npj Systems Biology and Applications*, 2:16015, August 2016.
- [44] Qiaonan Duan, St Patrick Reid, Neil R Clark, Zichen Wang, Nicolas F Fernandez, Andrew D Rouillard, Ben Readhead, Sarah R Tritsch, Rachel Hodos, Marc Hafner, Mario Niepel, Peter K Sorger, Joel T Dudley, Sina Bavari, Rekha G Panchal, and Avi Ma’ayan. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *npj Systems Biology and Applications*, 2:16015, 2016.
- [45] Yuzhu Duan, Daniel S Evans, Richard A Miller, Nicholas J Schork, Steven R Cummings, and Thomas Girke. signaturesearch: environment for gene expression signature searching and functional interpretation. *Nucleic Acids Res.*, October 2020.
- [46] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *Ann. Appl. Stat.*, 1(1):107–129, 2007.
- [47] Oana M Enache, David L Lahr, Ted E Natoli, Lev Litichevskiy, David Wadden, Corey Flynn, Joshua Gould, Jacob K Asiedu, Rajiv Narayan, and Aravind Subramanian. The GCTx format and `cmap{Py, R, M, J}` packages: resources for optimized storage and integrated traversal of annotated dense matrices. *Bioinformatics*, 35(8):1427–1429, April 2019.
- [48] Adnan Erol. The functions of PPARs in aging and longevity. *PPAR Res.*, 2007:39654, 2007.
- [49] Zhaoyuan Fang, Weidong Tian, and Hongbin Ji. A network-based gene-weighting approach for pathway analysis. *Cell Res.*, 22(3):565–580, March 2012.
- [50] Christian Feller and Ruedi Aebersold. A proteomic connectivity map. *Cell Syst*, 6(4):403–405, April 2018.
- [51] Zhaohui Feng and Arnold J Levine. The regulation of energy metabolism and the IGF-1/mTOR pathways by the p53 protein. *Trends Cell Biol.*, 20(7):427–434, July 2010.
- [52] Zhaohui Feng, Meihua Lin, and Rui Wu. The regulation of aging and longevity: A new and complex role of p53. *Genes Cancer*, 2(4):443–452, April 2011.

- [53] Luigi Fontana and Linda Partridge. Promoting health and longevity through diet: from model organisms to humans. *Cell*, 161(1):106–118, March 2015.
- [54] J Fort. Celecoxib, a COX-2-specific inhibitor: the clinical data. *Am. J. Orthop.*, 28(3 Suppl):13–18, March 1999.
- [55] Lazaros C Foukas, Benoit Bilanges, Lucia Bettedi, Wayne Pearce, Khaled Ali, Sara Sancho, Dominic J Withers, and Bart Vanhaesebroeck. Long-term p110 α PI3K inactivation exerts a beneficial effect on metabolism. *EMBO Mol. Med.*, 5(4):563–571, April 2013.
- [56] David Freedman, Robert Pisani, and Roger Purves. *Statistics, 4th Edition*. W. W. Norton & Company, 4 edition, February 2007.
- [57] Matías Fuentealba, Handan Melike Dönertaş, Rhianna Williams, Johnathan Labbadia, Janet M Thornton, and Linda Partridge. Using the drug-protein interactome to identify anti-ageing compounds for humans, 2019.
- [58] Matías Fuentealba, Handan Melike Dönertaş, Rhianna Williams, Johnathan Labbadia, Janet M Thornton, and Linda Partridge. Using the drug-protein interactome to identify anti-ageing compounds for humans. *PLoS Comput. Biol.*, 15(1):e1006639, January 2019.
- [59] Brigitte Ganter, Stuart Tugendreich, Cecelia I Pearson, Eser Ayanoglu, Susanne Baumhueter, Keith A Bostian, Lindsay Brady, Leslie J Browne, John T Calvin, Gwo-Jen Day, Naiomi Breckenridge, Shane Dunlea, Barrett P Eynon, L Mike Furness, Joe Ferng, Mark R Fielden, Susan Y Fujimoto, Li Gong, Christopher Hu, Radha Idury, Michael S B Judo, Kyle L Kolaja, May D Lee, Christopher McSorley, James M Minor, Ramesh V Nair, Georges Natsoulis, Peter Nguyen, Simone M Nicholson, Hang Pham, Alan H Roter, Dongxu Sun, Siqi Tan, Silke Thode, Alexander M Tolley, Antoaneta Vladimirova, Jian Yang, Zhiming Zhou, and Kurt Jarnagin. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.*, 119(3):219–244, September 2005.
- [60] Juan M García-Martínez, Jennifer Moran, Rosemary G Clarke, Alex Gray, Sabina C Cosulich, Christine M Chresta, and Dario R Alessi. Ku-0063794 is a specific inhibitor of the mammalian target of rapamycin (mTOR). *Biochem. J*, 421(1):29–42, June 2009.
- [61] Anna Gaulton, Anne Hersey, Michał Nowotka, A Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P Overington, George Papadatos, Ines Smit, and Andrew R Leach. The ChEMBL database in 2017. *Nucleic Acids Res.*, 45(D1):D945–D954, November 2016.
- [62] Marc Gillespie, Bijay Jassal, and Guanming Wu. Reactome analysis tools. <https://reactome.org/PathwayBrowser/#TOOL=AT>. Accessed: 2022-02-20.

- [63] M M Goldenberg. Celecoxib, a selective cyclooxygenase-2 inhibitor for the treatment of rheumatoid arthritis and osteoarthritis. *Clin. Ther.*, 21(9):1497–513; discussion 1427–8, September 1999.
- [64] Z G Goldsmith and D N Dhanasekaran. G protein regulation of MAPK networks. *Oncogene*, 26(22):3122–3142, May 2007.
- [65] Lata T Gooljarsingh, Christine Fernandes, Kang Yan, Hong Zhang, Michael Grooms, Kyung Johanson, Robert H Sinnamon, Robert B Kirkpatrick, John Kerrigan, Tia Lewis, Marc Arnone, Alastair J King, Zhihong Lai, Robert A Copeland, and Peter J Tummino. A biochemical rationale for the anticancer effects of hsp90 inhibitors: Slow, tight binding inhibition by geldanamycin and its analogues. *Proc. Natl. Acad. Sci. U. S. A.*, 103(20):7625–7630, May 2006.
- [66] Magdalena Gorska, Urszula Popowska, Alicja Sielicka-Dudzin, Alicja Kuban-Jankowska, Wojciech Sawczuk, Narcyz Knap, Giuseppe Cicero, and Fabio Wozniak. Geldanamycin and its derivatives as hsp90 inhibitors. *Front. Biosci.*, 17:2269–2277, June 2012.
- [67] Arvin M Gouw, Gizem Efe, Rita Barakat, Andrew Preecha, Morvarid Mehdizadeh, Steven A Garan, and George A Brooks. Roles of estrogen receptor-alpha in mediating life span: the hypothalamic deregulation hypothesis. *Physiol. Genomics*, 49(2):88–95, February 2017.
- [68] Graham J. G. Upton. Fisher’s exact test. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 155(3):395–402, 1992.
- [69] GTEx Consortium. The Genotype-Tissue expression (GTEx) project. *Nat. Genet.*, 45(6):580–585, June 2013.
- [70] L Guarente and C Kenyon. Genetic pathways that regulate ageing in model organisms. *Nature*, 408(6809):255–262, November 2000.
- [71] Artem P Gureev, Ekaterina A Shaforostova, and Vasily N Popov. Regulation of mitochondrial biogenesis as a way for active longevity: Interaction between the nrf2 and PGC-1 α signaling pathways. *Front. Genet.*, 10:435, May 2019.
- [72] Tyler W H Backman and Thomas Girke. systemPipeR: NGS workflow and report generation environment. *BMC Bioinformatics*, 17:388, September 2016.
- [73] Damir Hamamdžić, Robert S Fenning, Dhavalkumar Patel, Emile R Mohler, 3rd, Ksenia A Orlova, Alexander C Wright, Raul Llano, Martin G Keane, Richard P Shannon, Morris J Birnbaum, and Robert L Wilensky. Akt pathway is hypoactivated by synergistic actions of diabetes mellitus and hypercholesterolemia resulting in advanced coronary artery disease. *Am. J. Physiol. Heart Circ. Physiol.*, 299(3):H699–706, September 2010.

- [74] J A Hanley and B J McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, September 1983.
- [75] Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, Srinivas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Rob Deconde, Menzies Chen, Indika Rajapakse, Stephen Friend, Trey Ideker, and Kang Zhang. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell*, 49(2):359–367, January 2013.
- [76] David E Harrison, Randy Strong, Silvestre Alavez, Clinton Michael Astle, John DiGiovanni, Elizabeth Fernandez, Kevin Flurkey, Michael Garratt, Jonathan A L Gelfond, Martin A Javors, Moshe Levi, Gordon J Lithgow, Francesca Macchiarini, James F Nelson, Stacey J Sukoff Rizzo, Thomas J Slaga, Tim Stearns, John Erby Wilkinson, and Richard A Miller. Acarbose improves health and lifespan in aging HET3 mice. *Aging Cell*, 18(2):e12898, April 2019.
- [77] David E Harrison, Randy Strong, Silvestre Alavez, Clinton Michael Astle, John DiGiovanni, Elizabeth Fernandez, Kevin Flurkey, Michael Garratt, Jonathan A L Gelfond, Martin A Javors, Moshe Levi, Gordon J Lithgow, Francesca Macchiarini, James F Nelson, Stacey J Sukoff Rizzo, Thomas J Slaga, Tim Stearns, John Erby Wilkinson, and Richard A Miller. Acarbose improves health and lifespan in aging HET3 mice. *Aging Cell*, 18(2):e12898, April 2019.
- [78] David E Harrison, Randy Strong, Peter Reifsnnyder, Navasuja Kumar, Elizabeth Fernandez, Kevin Flurkey, Martin A Javors, Marisa Lopez-Cruzan, Francesca Macchiarini, James F Nelson, Alessandro Bitto, Amy L Sindler, Gino Cortopassi, Kylie Kavanagh, Lin Leng, Richard Bucala, Nadia Rosenthal, Adam Salmon, Timothy M Stearns, Molly Bogue, and Richard A Miller. 17- α -estradiol late in life extends lifespan in aging UM-HET3 male mice; nicotinamide riboside and three other drugs do not affect lifespan in either sex. *Aging Cell*, 20(5):e13328, May 2021.
- [79] Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, 44(D1):D1214–9, January 2016.
- [80] Katie L Hector, Malgorzata Lagisz, and Shinichi Nakagawa. The effect of resveratrol on longevity across species: a meta-analysis, 2012.
- [81] Susan J Hewlings and Douglas S Kalman. Curcumin: A review of its effects on human health. *Foods*, 6(10), October 2017.
- [82] Martin Holzenberger, Joëlle Dupont, Bertrand Ducos, Patricia Leneuve, Alain Gëloën, Patrick C Even, Pascale Cervera, and Yves Le Bouc. IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. *Nature*, 421(6919):182–187, January 2003.

- [83] Hui Hua, Qingbin Kong, Hongying Zhang, Jiao Wang, Ting Luo, and Yangfu Jiang. Targeting mTOR for cancer therapy. *J. Hematol. Oncol.*, 12(1):71, July 2019.
- [84] Xinyan Huang, Raelene A Charbeneau, Ying Fu, Kuljeet Kaur, Isabelle Gerin, Ormond A MacDougald, and Richard R Neubig. Resistance to diet-induced obesity and improved insulin sensitivity in mice with a regulator of G protein signaling-insensitive G184S gnai2 allele. *Diabetes*, 57(1):77–85, January 2008.
- [85] Michael W Hughes, Ting-Xin Jiang, Sung-Jan Lin, Yvonne Leung, Krzysztof Kobiela, Randall B Widelitz, and Cheng M Chuong. Disrupted ectodermal organ morphogenesis in mice with a conditional histone deacetylase 1, 2 deletion in the epidermis. *J. Invest. Dermatol.*, 134(1):24–32, January 2014.
- [86] T R Hughes, M J Marton, A R Jones, C J Roberts, R Stoughton, C D Armour, H A Bennett, E Coffey, H Dai, Y D He, M J Kidd, A M King, M R Meyer, D Slade, P Y Lum, S B Stepaniants, D D Shoemaker, D Gachotte, K Chakraburtt, J Simon, M Bard, and S H Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, July 2000.
- [87] Donald K Ingram, Min Zhu, Jacek Mameczarz, Sige Zou, Mark A Lane, George S Roth, and Rafael deCabo. Calorie restriction mimetics: an emerging research field. *Aging Cell*, 5(2):97–108, April 2006.
- [88] Francesco Iorio, Roberta Bosotti, Emanuela Scacheri, Vincenzo Belcastro, Pratibha Mithbaokar, Rosa Ferriero, Loredana Murino, Roberto Tagliaferri, Nicola Brunetti-Pierr, Antonella Isacchi, and Diego di Bernardo. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U. S. A.*, 107(33):14621–14626, August 2010.
- [89] Ulaş Işıldak, Mehmet Somel, Janet M Thornton, and Handan Melike Dönertaş. Temporal changes in the gene expression heterogeneity during brain development and aging. *Sci. Rep.*, 10(1):1–15, March 2020.
- [90] Eva Istvan. Statin inhibition of HMG-CoA reductase: a 3-dimensional view. *Atheroscler. Suppl.*, 4(1):3–8, March 2003.
- [91] Haomiao Jia, Matthew M Zack, William W Thompson, Alex E Crosby, and Irving I Gottesman. Impact of depression on quality-adjusted life expectancy (QALE) directly as well as indirectly through suicide. *Soc. Psychiatry Psychiatr. Epidemiol.*, 50(6):939–949, June 2015.
- [92] Simon C Johnson, Peter S Rabinovitch, and Matt Kaeberlein. mTOR is a key modulator of ageing and age-related disease. *Nature*, 493(7432):338–345, January 2013.
- [93] Simon C Johnson, Melana E Yanos, Ernst-Bernhard Kayser, Albert Quintana, Maya Sangesland, Anthony Castanza, Lauren Uhde, Jessica Hui, Valerie Z Wall, Arni Gagmidze, Kelly Oh, Brian M Wasko, Fresnida J Ramos, Richard D Palmiter, Peter S Rabinovitch, Philip G Morgan, Margaret M Sedensky, and Matt Kaeberlein. mTOR

- inhibition alleviates mitochondrial disease in a mouse model of leigh syndrome. *Science*, 342(6165):1524–1528, December 2013.
- [94] Meaghan J Jones, Sarah J Goodman, and Michael S Kobor. DNA methylation and healthy human aging. *Aging Cell*, 14(6):924–932, December 2015.
- [95] Riia K Junnila, Edward O List, Darlene E Berryman, John W Murrey, and John J Kopchick. The GH/IGF-1 axis in ageing and longevity. *Nat. Rev. Endocrinol.*, 9(6):366–376, June 2013.
- [96] Mahita Kadmiel and John A Cidlowski. Glucocorticoid receptor signaling in health and disease. *Trends Pharmacol. Sci.*, 34(9):518–530, September 2013.
- [97] Alexandra B Keenan, Sherry L Jenkins, Kathleen M Jagodnik, Simon Koplev, Edward He, Denis Torre, Zichen Wang, Anders B Dohlman, Moshe C Silverstein, Alexander Lachmann, Maxim V Kuleshov, Avi Ma’ayan, Vasileios Stathias, Raymond Terryn, Daniel Cooper, Michele Forlin, Amar Koleti, Dusica Vidovic, Caty Chung, Stephan C Schürer, Jouzas Vasiliauskas, Marcin Pilarczyk, Behrouz Shamsaei, Mehdi Fazel, Yan Ren, Wen Niu, Nicholas A Clark, Shana White, Naim Mahi, Lixia Zhang, Michal Kouril, John F Reichard, Siva Sivaganesan, Mario Medvedovic, Jaroslaw Meller, Rick J Koch, Marc R Birtwistle, Ravi Iyengar, Eric A Sobie, Evren U Azeloglu, Julia Kaye, Jeannette Osterloh, Kelly Haston, Jaslin Kalra, Steve Finkbiener, Jonathan Li, Pamela Milani, Miriam Adam, Renan Escalante-Chong, Karen Sachs, Alex Lenail, Divya Ramamoorthy, Ernest Fraenkel, Gavin Daigle, Uzma Hussain, Alyssa Coye, Jeffrey Rothstein, Dhruv Sareen, Loren Ornelas, Maria Banuelos, Berhan Mandefro, Ritchie Ho, Clive N Svendsen, Ryan G Lim, Jennifer Stocksdale, Malcolm S Casale, Terri G Thompson, Jie Wu, Leslie M Thompson, Victoria Dardov, Vidya Venkatraman, Andrea Matlock, Jennifer E Van Eyk, Jacob D Jaffe, Malvina Papanastasiou, Aravind Subramanian, Todd R Golub, Sean D Erickson, Mohammad Fallahi-Sichani, Marc Hafner, Nathanael S Gray, Jia-Ren Lin, Caitlin E Mills, Jeremy L Muhlich, Mario Niepel, Caroline E Shamu, Elizabeth H Williams, David Wrobel, Peter K Sorger, Laura M Heiser, Joe W Gray, James E Korkola, Gordon B Mills, Mark LaBarge, Heidi S Feiler, Mark A Dane, Elmar Bucher, Michel Nederlof, Damir Sudar, Sean Gross, David F Kilburn, Rebecca Smith, Kaylyn Devlin, Ron Margolis, Leslie Derr, Albert Lee, and Ajay Pillai. The library of integrated Network-Based cellular signatures NIH program: System-Level cataloging of human cells response to perturbations. *Cell Syst*, 6(1):13–24, January 2018.
- [98] C Kenyon. A conserved regulatory system for aging. *Cell*, 105(2):165–168, April 2001.
- [99] Hyo Won Kim, Jiyoung Kim, Jaekyoon Kim, Siyoung Lee, Bo-Ryoung Choi, Jung-Soo Han, Ki Won Lee, and Hyong Joo Lee. 3,3-diindolylmethane inhibits Lipopolysaccharide-Induced microglial hyperactivation and attenuates brain inflammation. *Toxicol. Sci.*, 137(1):158–167, October 2013.
- [100] Hyun-Jung Kim and Suk-Chul Bae. Histone deacetylase inhibitors: molecular mechanisms of action and clinical trials as anti-cancer drugs. *Am. J. Transl. Res.*, 3(2):166–179, February 2011.

- [101] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, 49(D1):D1388–D1395, January 2021.
- [102] Viktoria Koroknai, Vikas Patel, István Szász, Róza Ádány, and Margit Balazs. Gene expression signature of BRAF inhibitor resistant melanoma spheroids. *Pathol. Oncol. Res.*, 26(4):2557–2566, October 2020.
- [103] J Kotsopoulos, S Zhang, M Akbari, L Salmena, M Llacuachaqui, M Zelig, P Sun, and S A Narod. BRCA1 mRNA levels following a 4–6-week intervention with oral 3,3-diindolylmethane. *Br. J. Cancer*, 111(7):1269–1274, July 2014.
- [104] Michael Kuhn, Damian Szklarczyk, Andrea Franceschini, Monica Campillos, Christian von Mering, Lars Juhl Jensen, Andreas Beyer, and Peer Bork. STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*, 38(Database issue):D552–6, January 2010.
- [105] Kei Kunimasa, Tatsuya Nagano, Yohei Shimono, Ryota Dokuni, Tatsunori Kiri, Shuntaro Tokunaga, Daisuke Tamura, Masatsugu Yamamoto, Motoko Tachihara, Kazuyuki Kobayashi, Miyako Satouchi, and Yoshihiro Nishimura. Glucose metabolism-targeted therapy and withaferin a are effective for epidermal growth factor receptor tyrosine kinase inhibitor-induced drug-tolerant persisters. *Cancer Sci.*, 108(7):1368–1377, July 2017.
- [106] Imge Kunter, Esra Erdal, Deniz Nart, Funda Yilmaz, Sedat Karademir, Ozgul Sagol, and Nese Atabey. Active form of AKT controls cell proliferation and response to apoptosis in hepatocellular carcinoma. *Oncol. Rep.*, 31(2):573–580, February 2014.
- [107] Liming Lai, Jason Hennessey, Valerie Bares, Eun Woo Son, Yuguang Ban, Wei Wang, Jianli Qi, Gaixin Jiang, Arthur Liberzon, and Steven Xijin Ge. GSKB: A gene set database for pathway analysis in mouse. October 2016.
- [108] Ranjani Lakshminarasimhan and Gangning Liang. The role of DNA methylation in cancer. *Adv. Exp. Med. Biol.*, 945:151–172, 2016.
- [109] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A Armstrong, Stephen J Haggarty, Paul A Clemons, Ru Wei, Steven A Carr, Eric S Lander, and Todd R Golub. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, September 2006.
- [110] Dudley W Lamming, Lan Ye, David M Sabatini, and Joseph A Baur. Rapalogs and mTOR inhibitors as anti-aging therapeutics. *J. Clin. Invest.*, 123(3):980–989, March 2013.
- [111] Mario Lauria. Rank-based transcriptional signatures. *Systems Biomedicine*, 1(4):228–239, October 2013.

- [112] Gang Jun Lee, Jin Ju Lim, and Seogang Hyun. Minocycline treatment increases resistance to oxidative stress and extends lifespan in drosophila via FOXO. *Oncotarget*, 8(50):87878–87890, October 2017.
- [113] R Y Lee, J Hench, and G Ruvkun. Regulation of *c. elegans* DAF-16 and its human ortholog FKHL1 by the *daf-2* insulin-like signaling pathway. *Curr. Biol.*, 11(24):1950–1957, December 2001.
- [114] Ellis R Levin. Extranuclear estrogen receptor’s roles in physiology: lessons from mouse models. *Am. J. Physiol. Endocrinol. Metab.*, 307(2):E133–40, July 2014.
- [115] Yufei Li, Nathaniel W Mahloch, Nicholas J E Starkey, Mónica Peña-Luna, George E Rottinghaus, Kevin L Fritsche, Cynthia Besch-Williford, and Dennis B Lubahn. 3,3-diindolylmethane Dose-Dependently prevents advanced prostate cancer. April 2019.
- [116] James K Liao, Minoru Seto, and Kensuke Noma. Rho kinase (ROCK) inhibitors. *J. Cardiovasc. Pharmacol.*, 50(1):17–24, July 2007.
- [117] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst*, 1(6):417–425, December 2015.
- [118] K Lin, H Hsin, N Libina, and C Kenyon. Regulation of the *caenorhabditis elegans* longevity protein DAF-16 by insulin/IGF-1 and germline signaling. *Nat. Genet.*, 28(2):139–145, June 2001.
- [119] Tsang-Pai Liu, Yao-Yu Hsieh, Chia-Jung Chou, and Pei-Ming Yang. Systematic polypharmacology and drug repurposing via an integrated 11000-based connectivity map database mining. *R. Soc. Open Sci.*, 5(11):181321, November 2018.
- [120] Yan Lu, Jianjun Chen, Min Xiao, Wei Li, and Duane D Miller. An overview of tubulin inhibitors that interact with the colchicine binding site. *Pharm. Res.*, 29(11):2943–2971, November 2012.
- [121] Thomas J Lynch, Daphne W Bell, Raffaella Sordella, Sarada Gurubhagavatula, Ross A Okimoto, Brian W Brannigan, Patricia L Harris, Sara M Haserlat, Jeffrey G Supko, Frank G Haluska, David N Louis, David C Christiansi, Jeff Settleman, and Daniel A Haber. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.*, 350(21):2129–2139, May 2004.
- [122] Marcello Maggio, Jack M Guralnik, Dan L Longo, and Luigi Ferrucci. Interleukin-6 in aging and chronic disease: a magnificent pathway. *J. Gerontol. A Biol. Sci. Med. Sci.*, 61(6):575–584, June 2006.
- [123] H R Massie, V R Aiello, T R Williams, M B Baird, and J L Hough. Effect of vitamin a on longevity. *Exp. Gerontol.*, 28(6):601–610, November 1993.

- [124] Caio Henrique Mazucanti, João Victor Cabral-Costa, Andrea Rodrigues Vasconcelos, Diana Zukas Andreotti, Cristoforo Scavone, and Elisa Mitiko Kawamoto. Longevity pathways (mTOR, SIRT, Insulin/IGF-1) as key modulatory targets on aging and neurodegeneration. *Curr. Top. Med. Chem.*, 15(21):2116–2138, 2015.
- [125] Amir Mehrgou and Mansoureh Akouchekian. The importance of BRCA1 and BRCA2 genes mutations in breast cancer development. *Med. J. Islam. Repub. Iran*, 30:369, May 2016.
- [126] Jennifer A Messing, Roschelle Heuberger, and Jennifer A Schisa. Effect of vitamin D3 on lifespan in *caenorhabditis elegans*. *Curr. Aging Sci.*, 6(3):220–224, December 2013.
- [127] Badri N Mishra. Secret of eternal youth; teaching from the centenarian hot spots (“blue zones”). *Indian J. Community Med.*, 34(4):273–275, October 2009.
- [128] Y Miyata. Hsp90 inhibitor geldanamycin and its derivatives as novel cancer chemotherapeutic agents. *Curr. Pharm. Des.*, 11(9):1131–1138, 2005.
- [129] Hideyuki Miyauchi, Tohru Minamino, Kaoru Tateno, Takeshige Kunieda, Haruhiro Toko, and Issei Komuro. Akt negatively regulates the in vitro lifespan of human endothelial cells via a p53/p21-dependent pathway. *EMBO J.*, 23(1):212–220, January 2004.
- [130] Ismail S Mohiuddin and Min H Kang. DNA-PK as an emerging therapeutic target in cancer. *Front. Oncol.*, 9:635, July 2019.
- [131] J Z Morris, H A Tissenbaum, and G Ruvkun. A phosphatidylinositol-3-OH kinase family member regulating longevity and diapause in *caenorhabditis elegans*. *Nature*, 382(6591):536–539, August 1996.
- [132] A A Moskalev and M V Shaposhnikov. Pharmacological inhibition of phosphoinositide 3 and TOR kinases improves survival of *drosophila melanogaster*. *Rejuvenation Res.*, 13(2-3):246–247, April 2010.
- [133] Farnaz Najmi Varzaneh, Farshad Sharifi, Arash Hossein-Nezhad, Mojde Mirarefin, Zhila Maghbooli, Maryam Ghaderpanahi, Bagher Larijani, and Hossein Fakhrzadeh. Association of vitamin D receptor with longevity and healthy aging. *Acta Med. Iran.*, 51(4):236–241, May 2013.
- [134] Tetsuya Okuyama, Hideki Inoue, Sadatsugu Ookuma, Takayuki Satoh, Kei Kano, Sakiko Honjoh, Naoki Hisamoto, Kunihiro Matsumoto, and Eisuke Nishida. The ERK-MAPK pathway regulates longevity through SKN-1 and insulin-like signaling in *caenorhabditis elegans*. *J. Biol. Chem.*, 285(39):30274–30281, September 2010.
- [135] Pál Pacher, Joseph S Beckman, and Lucas Liaudet. Nitric oxide and peroxynitrite in health and disease. *Physiol. Rev.*, 87(1):315–424, January 2007.

- [136] Clare Pacini, Francesco Iorio, Emanuel Gonçalves, Murat Iskar, Thomas Klabunde, Peer Bork, and Julio Saez-Rodriguez. DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics*, 29(1):132–134, January 2013.
- [137] Haihui Pan and Toren Finkel. Key proteins and pathways that regulate lifespan. *J. Biol. Chem.*, 292(16):6452–6460, April 2017.
- [138] David Papadopoli, Karine Boulay, Lawrence Kazak, Michael Pollak, Frédéric Mallette, Ivan Topisirovic, and Laura Hulea. mTOR as a central regulator of lifespan and aging. *F1000Res.*, 8, July 2019.
- [139] S Paradis and G Ruvkun. *Caenorhabditis elegans* Akt/PKB transduces insulin receptor-like signals from AGE-1 PI3 kinase to the DAF-16 transcription factor. *Genes Dev.*, 12(16):2488–2498, August 1998.
- [140] David Peck, Emily D Crawford, Kenneth N Ross, Kimberly Stegmaier, Todd R Golub, and Justin Lamb. A method for high-throughput gene expression signature analysis. *Genome Biol.*, 7(7):R61, 2006.
- [141] Sagit Peleg, Dalia Varon, Tatiana Ivanina, Carmen W Dessauer, and Nathan Dascal. G(α)(i) controls the gating of the G protein-activated k(+) channel, GIRK. *Neuron*, 33(1):87–99, January 2002.
- [142] Stuart D Pepper, Emma K Saunders, Laura E Edwards, Claire L Wilson, and Crispin J Miller. The utility of MAS5 expression summary and detection call algorithms. *BMC Bioinformatics*, 8:273, July 2007.
- [143] Thomas Perls, Iliana V Kohler, Stacy Andersen, Emily Schoenhofen, Jaemi Pennington, Robert Young, Dellara Terry, and Irma T Elo. Survival of parents and siblings of supercentenarians. *J. Gerontol. A Biol. Sci. Med. Sci.*, 62(9):1028–1034, September 2007.
- [144] Thomas T Perls, John Wilmoth, Robin Levenson, Maureen Drinkwater, Melissa Cohen, Hazel Bogan, Erin Joyce, Stephanie Brewster, Louis Kunkel, and Annibale Puca. Life-long sustained mortality advantage of siblings of centenarians. *Proc. Natl. Acad. Sci. U. S. A.*, 99(12):8442–8447, June 2002.
- [145] Marjolein J Peters, Roby Joehanes, Luke C Pilling, Claudia Schurmann, Karen N Conneely, Joseph Powell, Eva Reinmaa, George L Sutphin, Alexandra Zhernakova, Katharina Schramm, Yana A Wilson, Sayuko Kobes, Taru Tukiainen, NABEC/UK-BEC Consortium, Yolande F Ramos, Harald H H Göring, Myriam Fornage, Yongmei Liu, Sina A Gharib, Barbara E Stranger, Philip L De Jager, Abraham Aviv, Daniel Levy, Joanne M Murabito, Peter J Munson, Tianxiao Huan, Albert Hofman, André G Uitterlinden, Fernando Rivadeneira, Jeroen van Rooij, Lisette Stolk, Linda Broer, Michael M P J Verbiest, Mila Jhamai, Pascal Arp, Andres Metspalu, Liina Tserel, Lili Milani, Nilesh J Samani, Pärt Peterson, Silva Kasela, Veryan Codd, Annette Peters, Cavin K Ward-Caviness, Christian Herder, Melanie Waldenberger, Michael

- Roden, Paula Singmann, Sonja Zeilinger, Thomas Illig, Georg Homuth, Hans-Jürgen Grabe, Henry Völzke, Leif Steil, Thomas Kocher, Anna Murray, David Melzer, Hanieh Yaghootkar, Stefania Bandinelli, Eric K Moses, Jack W Kent, Joanne E Curran, Matthew P Johnson, Sarah Williams-Blangero, Harm-Jan Westra, Allan F McRae, Jennifer A Smith, Sharon L R Kardia, Iris Hovatta, Markus Perola, Samuli Ripatti, Veikko Salomaa, Anjali K Henders, Nicholas G Martin, Alicia K Smith, Divya Mehta, Elisabeth B Binder, K Maria Nylocks, Elizabeth M Kennedy, Torsten Klen-
 gel, Jingzhong Ding, Astrid M Suchy-Dicey, Daniel A Enquobahrie, Jennifer Brody, Jerome I Rotter, Yii-Der I Chen, Jeanine Houwing-Duistermaat, Margreet Kloppen-
 burg, P Eline Slagboom, Quinta Helmer, Wouter den Hollander, Shannon Bean, Tow-
 fique Raj, Noman Bakhshi, Qiao Ping Wang, Lisa J Oyston, Bruce M Psaty, Russell P
 Tracy, Grant W Montgomery, Stephen T Turner, John Blangero, Ingrid Meulenbelt,
 Kerry J Ressler, Jian Yang, Lude Franke, Johannes Kettunen, Peter M Visscher,
 G Gregory Neely, Ron Korstanje, Robert L Hanson, Holger Prokisch, Luigi Ferrucci,
 Tonu Esko, Alexander Teumer, Joyce B J van Meurs, and Andrew D Johnson. The
 transcriptional landscape of age in human peripheral blood. *Nat. Commun.*, 6:8570,
 October 2015.
- [146] Michael Petrascheck, Xiaolan Ye, and Linda B Buck. An antidepressant that extends
 lifespan in adult *Caenorhabditis elegans*. *Nature*, 450(7169):553–556, November 2007.
- [147] Marcin Pilarczyk, Mehdi Fazel Najafabadi, Michal Kouril, Juozas Vasiliauskas, Wen
 Niu, Behrouz Shamsaei, Naim Mahi, Lixia Zhang, Nicholas Clark, Yan Ren, Shana
 White, Rashid Karim, Huan Xu, Jacek Biesiada, Mark F Bennet, Sarah Davidson,
 John F Reichard, Vasileios Stathias, Amar Koleti, Dusica Vidovic, Daniel J B Clark,
 Stephan Schurer, Avi Ma’ayan, Jarek Meller, and Mario Medvedovic. Connecting
 omics signatures of diseases, drugs, and mechanisms of actions with iLINCS. October
 2019.
- [148] Julie M Pinkston, Delia Garigan, Malene Hansen, and Cynthia Kenyon. Mutations
 that increase the life span of *C. elegans* inhibit tumor growth. *Science*, 313(5789):971–
 975, August 2006.
- [149] Andrea Princz, Federico Pelisch, and Nektarios Tavernarakis. SUMO promotes
 longevity and maintains mitochondrial homeostasis during ageing in *Caenorhabditis*
elegans. *Sci. Rep.*, 10(1):15513, September 2020.
- [150] Thandla Raghavendra. Neuromuscular blocking drugs: discovery and development. *J.*
R. Soc. Med., 95(7):363–367, July 2002.
- [151] Murugesan V S Rajaram, Latha P Ganesan, Kishore V L Parsa, Jonathan P
 Butchar, John S Gunn, and Susheela Tridandapani. Akt/Protein kinase B modu-
 lates macrophage inflammatory response to francisella infection and confers a survival
 advantage in mice. *J. Immunol.*, 177(9):6317–6324, November 2006.
- [152] Sunitha Rangaraju, Gregory M Solis, Sofia I Andersson, Rafael L Gomez-Amaro,
 Rozina Kardakaris, Caroline D Broaddus, Alexander B Niculescu, 3rd, and Michael

- Petrascheck. Atypical antidepressants extend lifespan of *Caenorhabditis elegans* by activation of a non-cell-autonomous stress response. *Aging Cell*, 14(6):971–981, December 2015.
- [153] Irene Maeve Rea, David S Gibson, Victoria McGilligan, Susan E McNerlan, H Denis Alexander, and Owen A Ross. Age and Age-Related diseases: Role of inflammation triggers and cytokines. *Front. Immunol.*, 9:586, April 2018.
- [154] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47–e47, January 2015.
- [155] Nilsa Rivera-Del Valle, Shan Gao, Claudia P Miller, Joy Fulbright, Carolina Gonzales, Mint Sirisawad, Susanne Steggerda, Jennifer Wheeler, Sriram Balasubramanian, and Joya Chandra. PCI-24781, a novel hydroxamic acid HDAC inhibitor, exerts cytotoxicity and histone alterations via caspase-8 and FADD in leukemia cells. *Int. J. Cell Biol.*, 2010:207420, January 2010.
- [156] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. pROC: an open-source package for R and s+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77, March 2011.
- [157] S Mark Roe, Chrisostomos Prodromou, Ronan O’Brien, John E Ladbury, Peter W Piper, and Laurence H Pearl. Structural basis for inhibition of the hsp90 molecular chaperone by the antitumor antibiotics radicicol and geldanamycin. *J. Med. Chem.*, 42(2):260–266, January 1999.
- [158] Irene Gallego Romero, Ilya Ruvinsky, and Yoav Gilad. Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.*, 13(7):505–516, June 2012.
- [159] Christopher Rongo. Epidermal growth factor and aging: a signaling molecule reveals a new eye opening function. *Aging*, 3(9):896–905, September 2011.
- [160] Owen A Ross, Martin D Curran, Karen A Crum, I Maeve Rea, Yvonne A Barnett, and Derek Middleton. Increased frequency of the 2437T allele of the heat shock protein 70-hom gene in an aged Irish population. *Exp. Gerontol.*, 38(5):561–565, May 2003.
- [161] Antero Salminen and Kai Kaarniranta. AMP-activated protein kinase (AMPK) controls the aging process via an integrated signaling network. *Ageing Res. Rev.*, 11(2):230–241, April 2012.
- [162] Thomas Sandmann, Sarah K Kummerfeld, Robert Gentleman, and Richard Bourgon. gCMAP: user-friendly connectivity mapping with R. *Bioinformatics*, 30(1):127–128, January 2014.
- [163] F Schächter, L Faure-Delanef, F Guénot, H Rouger, P Froguel, L Lesueur-Ginot, and D Cohen. Genetic associations with human longevity at the APOE and ACE loci. *Nat. Genet.*, 6(1):29–32, January 1994.

- [164] Manja Schoenmaker, Anton J M de Craen, Paul H E M de Meijer, Marian Beekman, Gerard J Blauw, P Eline Slagboom, and Rudi G J Westendorp. Evidence of genetic enrichment for exceptional survival using a family approach: the leiden longevity study. *Eur. J. Hum. Genet.*, 14(1):79–84, January 2006.
- [165] Sebastiano Sciarretta, Maurizio Forte, Francesca Castoldi, Giacomo Frati, Francesco Versaci, Junichi Sadoshima, Guido Kroemer, and Maria Chiara Maiuri. Caloric restriction mimetics for the treatment of cardiovascular diseases. *Cardiovasc. Res.*, 117(6):1434–1449, October 2020.
- [166] Kathleen Petri Seiler, Gregory A George, Mary Pat Happ, Nicole E Bodycombe, Hyman A Carrinski, Stephanie Norton, Steve Brudz, John P Sullivan, Jeremy Muhllich, Martin Serrano, Paul Ferraiolo, Nicola J Tolliday, Stuart L Schreiber, and Paul A Clemons. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, 36(Database issue):D351–9, January 2008.
- [167] Xiaoyin Shan, Cleresa Roberts, Yemin Lan, and Ivona Percec. Age Alters Chromatin Structure and Expression of SUMO Proteins under Stress Conditions in Human Adipose-Derived Stem Cells. *Sci. Rep.*, 8(1):11502, July 2018.
- [168] Y Shao, T He, G J Fisher, J J Voorhees, and T Quan. Molecular basis of retinol anti-ageing properties in naturally aged human skin in vivo. *Int. J. Cosmet. Sci.*, 39(1):56–65, February 2017.
- [169] J N Sharma, A Al-Omran, and S S Parvathy. Role of nitric oxide in inflammatory diseases. *Inflammopharmacology*, 15(6):252–259, December 2007.
- [170] Wendy M Shaw, Shijing Luo, Jessica Landis, Jasmine Ashraf, and Coleen T Murphy. The *c. elegans* TGF-beta dauer pathway regulates longevity via insulin signaling. *Curr. Biol.*, 17(19):1635–1645, October 2007.
- [171] Li-Rong Shen, Laurence D Parnell, Jose M Ordovas, and Chao-Qiang Lai. Curcumin and aging. *Biofactors*, 39(1):133–140, January 2013.
- [172] Giselle Saulnier Sholler, Erika A Currier, Akshita Dutta, Marni A Slavik, Sharon A Illenye, Maria Cecilia F Mendonca, Julie Dragon, Stephen S Roberts, and Jeffrey P Bond. PCI-24781 (abexinostat), a novel histone deacetylase inhibitor, induces reactive oxygen species-dependent apoptosis and is synergistic with bortezomib in neuroblastoma. *J. Cancer Res. Ther.*, 2:21, December 2013.
- [173] John C Siavelis, Marilena M Bourdakou, Emmanouil I Athanasiadis, George M Spyrou, and Konstantina S Nikita. Bioinformatics methods in drug repurposing for alzheimer’s disease. *Brief. Bioinform.*, 17(2):322–335, March 2016.
- [174] Smithamol Sithara, Tamsyn M Crowley, Ken Walder, and Kathryn Aston-Mourney. Gene expression signature: a powerful approach for drug discovery in diabetes. *J. Endocrinol.*, 232(2):R131–R139, February 2017.

- [175] Cathy Slack. Ras signaling in aging and metabolic regulation. *Nutr Healthy Aging*, 4(3):195–205, December 2017.
- [176] Cathy Slack, Nazif Alic, Andrea Foley, Melissa Cabecinha, Matthew P Hoddinott, and Linda Partridge. The Ras-Erk-ETS-Signaling pathway is a drug target for longevity. *Cell*, 162(1):72–83, July 2015.
- [177] Daniel L Smith, Jr, Rachael M Orlandella, David B Allison, and Lyse A Norian. Diabetes medications as potential calorie restriction mimetics—a focus on the alpha-glucosidase inhibitor acarbose. *Geroscience*, 43(3):1123–1133, June 2021.
- [178] Gregory M Solis, Rozina Kardakar, Elizabeth R Valentine, Liron Bar-Peled, Alice L Chen, Megan M Blewett, Mark A McCormick, James R Williamson, Brian Kennedy, Benjamin F Cravatt, and Michael Petrascheck. Translation attenuation by minocycline enhances longevity and proteostasis in old post-stress-responsive organisms. *Elife*, 7, November 2018.
- [179] Dylan C Souder and Rozalyn M Anderson. An expanding GSK3 network: implications for aging research. *Geroscience*, 41(4):369–382, August 2019.
- [180] Andreea L Stancu. AMPK activation can delay aging. *Discoveries (Craiova)*, 3(4):e53, December 2015.
- [181] Vasileios Stathias, John Turner, Amar Koleti, Dusica Vidovic, Daniel Cooper, Mehdi Fazel-Najafabadi, Marcin Pilarczyk, Raymond Terryn, Caty Chung, Afoma Umeano, Daniel J B Clarke, Alexander Lachmann, John Erol Evangelista, Avi Ma’ayan, Mario Medvedovic, and Stephan C Schürer. LINCS data portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Res.*, 48(D1):D431–D439, November 2019.
- [182] Vasileios Stathias, John Turner, Amar Koleti, Dusica Vidovic, Daniel Cooper, Mehdi Fazel-Najafabadi, Marcin Pilarczyk, Raymond Terryn, Caty Chung, Afoma Umeano, Daniel J B Clarke, Alexander Lachmann, John Erol Evangelista, Avi Ma’ayan, Mario Medvedovic, and Stephan C Schürer. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Res.*, 48:D431–D439, 2020.
- [183] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, April 2009.
- [184] Randy Strong, Richard A Miller, Adam Antebi, Clinton M Astle, Molly Bogue, Martin S Denzel, Elizabeth Fernandez, Kevin Flurkey, Karyn L Hamilton, Dudley W Lamming, Martin A Javors, João Pedro de Magalhães, Paul Anthony Martinez, Joe M McCord, Benjamin F Miller, Michael Müller, James F Nelson, Juliet Ndukum, G Ed Rainger, Arlan Richardson, David M Sabatini, Adam B Salmon, James W Simpkins, Wilma T Steegenga, Nancy L Nadon, and David E Harrison. Longer lifespan in male mice treated with a weakly estrogenic agonist, an antioxidant, an α -glucosidase inhibitor or a nrf2-inducer. *Aging Cell*, 15(5):872–884, October 2016.

- [185] Randy Strong, Richard A Miller, Clinton M Astle, Robert A Floyd, Kevin Flurkey, Kenneth L Hensley, Martin A Javors, Christiaan Leeuwenburgh, James F Nelson, Ennio Ongini, Nancy L Nadon, Huber R Warner, and David E Harrison. Nordihydroguaiaretic acid and aspirin increase lifespan of genetically heterogeneous male mice. *Aging Cell*, 7(5):641–650, October 2008.
- [186] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, David L Lahr, Jodi E Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M Enache, Federica Piccioni, Sarah A Johnson, Nicholas J Lyons, Alice H Berger, Alykhan F Shamji, Angela N Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S Gray, Paul A Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J Haggarty, Lucienne V Ronco, Jesse S Boehm, Stuart L Schreiber, John G Doench, Joshua A Bittker, David E Root, Bang Wong, and Todd R Golub. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, November 2017.
- [187] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, October 2005.
- [188] Damian Szklarczyk, Alberto Santos, Christian von Mering, Lars Juhl Jensen, Peer Bork, and Michael Kuhn. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.*, 44(D1):D380–4, January 2016.
- [189] Samira Tabaei and Seyyedeh Samaneh Tabae. DNA methylation abnormalities in atherosclerosis. *Artif. Cells Nanomed. Biotechnol.*, 47(1):2031–2041, December 2019.
- [190] Nicole M Templeman and Coleen T Murphy. Regulation of reproduction and longevity by nutrient-sensing pathways. *J. Cell Biol.*, 217(1):93–106, January 2018.
- [191] Dellara F Terry, Marsha A Wilcox, Maegan A McCormick, Jaemi Y Pennington, Emily A Schoenhofen, Stacy L Andersen, and Thomas T Perls. Lower all-cause, cardiovascular, and cancer mortality in centenarians’ offspring. *J. Am. Geriatr. Soc.*, 52(12):2074–2076, December 2004.
- [192] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, 45(D1):D158–D169, January 2017.
- [193] Xiaolin Tian. Dopaminergic neurons regulate aging and longevity in flies. June 2020.
- [194] Stuart D Tyner, Sundaresan Venkatachalam, Jene Choi, Stephen Jones, Nader Ghebranious, Herbert Igelmann, Xiongbin Lu, Gabrielle Soron, Benjamin Cooper, Cory

- Brayton, Sang Hee Park, Timothy Thompson, Gerard Karsenty, Allan Bradley, and Lawrence A Donehower. p53 mutant mice that display early ageing-associated phenotypes. *Nature*, 415(6867):45–53, January 2002.
- [195] Alexander Tyshkovskiy, Perinur Bozaykut, Anastasia A Borodinova, Maxim V Gerashchenko, Gene P Ables, Michael Garratt, Philipp Khaitovich, Clary B Clish, Richard A Miller, and Vadim N Gladyshev. Identification and application of gene expression signatures associated with lifespan extension. *Cell Metab.*, 30(3):573–593.e8, September 2019.
- [196] Matthew Ulgherait, Anil Rana, Michael Rera, Jacqueline Graniel, and David W Walker. AMPK modulates tissue and organismal aging in a non-cell-autonomous manner. *Cell Rep.*, 8(6):1767–1780, September 2014.
- [197] Diana van Heemst, Simon P Mooijaart, Marian Beekman, Jeroen Schreuder, Anton J M de Craen, Bernd W Brandt, P Eline Slagboom, Rudi G J Westendorp, and Long Life study group. Variation in the human TP53 gene affects old age survival and cancer mortality. *Exp. Gerontol.*, 40(1-2):11–15, January 2005.
- [198] Miguel Vazquez, Ruben Nogales-Cadenas, Javier Arroyo, Pedro Botias, Raul Garcia, Jose M Carazo, Francisco Tirado, Alberto Pascual-Montano, and Pedro Carmona-Saez. MARQ: an online tool to mine GEO for experiments with similar or opposite gene expression signatures. *Nucleic Acids Res.*, 38(suppl_2):W228–W232, May 2010.
- [199] Natascia Ventura, Shane L Rea, Alfonso Schiavi, Alessandro Torgovnick, Roberto Testi, and Thomas E Johnson. p53/CEP-1 increases or decreases lifespan, depending on level of mitochondrial bioenergetic stress. *Aging Cell*, 8(4):380–393, August 2009.
- [200] Alan S Verkman and Luis J V Galletta. Chloride channels as drug targets. *Nat. Rev. Drug Discov.*, 8(2):153–171, February 2009.
- [201] Ana B Villaseñor-Altamirano, Marco Moretto, Mariel Maldonado, Alejandra Zayas-Del Moral, Adrián Munguía-Reyes, Yair Romero, Jair S García-Sotelo, Luis A Aguilar, Oscar Aldana-Assad, Kristof Engelen, Moisés Selman, Julio Collado-Vides, Yalbi I Balderas-Martínez, and Alejandra Medina-Rivera. PulmonDB: a curated lung disease gene expression database. *Sci. Rep.*, 10(1):1–9, January 2020.
- [202] Mark Wade, Yunyuan V Wang, and Geoffrey M Wahl. The p53 orchestra: Mdm2 and mdmx set the tone. *Trends Cell Biol.*, 20(5):299–309, May 2010.
- [203] Yan Wang, Tyler W H Backman, Kevin Horan, and Thomas Girke. fmcsr: mismatch tolerant maximum common substructure searching in R. *Bioinformatics*, 29(21):2792–2794, November 2013.
- [204] Zhixiang Wang. ErbB receptors and cancer. *Methods Mol. Biol.*, 1652:3–35, 2017.
- [205] Zichen Wang, Caroline D Monteiro, Kathleen M Jagodnik, Nicolas F Fernandez, Gregory W Gundersen, Andrew D Rouillard, Sherry L Jenkins, Axel S Feldmann, Kevin S

- Hu, Michael G McDermott, Qiaonan Duan, Neil R Clark, Matthew R Jones, Yan Kou, Troy Goff, Holly Woodland, Fabio M R Amaral, Gregory L Szeto, Oliver Fuchs, Sophia M Schüssler-Fiorenza Rose, Shvetank Sharma, Uwe Schwartz, Xabier Bengoetxea Bausela, Maciej Szymkiewicz, Vasileios Maroulis, Anton Salykin, Carolina M Barra, Candice D Kruth, Nicholas J Bongio, Vaibhav Mathur, Radmila D Todoric, Udi E Rubin, Apostolos Malatras, Carl T Fulp, John A Galindo, Ruta Motiejunaite, Christoph Jüschke, Philip C Dishuck, Katharina Lahl, Mohieddin Jafari, Sara Aibar, Apostolos Zaravinos, Linda H Steenhuizen, Lindsey R Allison, Pablo Gamallo, Fernando de Andres Segura, Tyler Dae Devlin, Vicente Pérez-García, and Avi Ma'ayan. Extraction and analysis of signatures from the gene expression omnibus by the crowd. *Nat. Commun.*, 7(1):1–11, September 2016.
- [206] Huber R Warner. NIA's intervention testing program at 10 years of age. *Age*, 37(2):22, March 2015.
- [207] Ping Wee and Zhixiang Wang. Epidermal growth factor receptor cell proliferation signaling pathways. *Cancers*, 9(5), May 2017.
- [208] Thomas Weichhart. mTOR as regulator of lifespan, aging, and cellular senescence: A Mini-Review. *Gerontology*, 64(2):127–134, 2018.
- [209] Rudi G J Westendorp, Diana van Heemst, Maarten P Rozing, Marijke Frölich, Simon P Mooijaart, Gerard-Jan Blauw, Marian Beekman, Bastiaan T Heijmans, Anton J M de Craen, P Eline Slagboom, and Leiden Longevity Study Group. Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The leiden longevity study. *J. Am. Geriatr. Soc.*, 57(9):1634–1637, September 2009.
- [210] Nina Wettschureck and Stefan Offermanns. Mammalian G proteins and their cell type specific functions. *Physiol. Rev.*, 85(4):1159–1204, October 2005.
- [211] Morris F White. Longevity: Mapping the path to a longer life. *Nature*, 524(7564):170–171, August 2015.
- [212] M J Wieduwilt and M M Moasser. The epidermal growth factor receptor family: biology driving targeted therapeutics. *Cell. Mol. Life Sci.*, 65(10):1566–1584, May 2008.
- [213] Bradley J Willcox, Timothy A Donlon, Qimei He, Randi Chen, John S Grove, Katsuhiko Yano, Kamal H Masaki, D Craig Willcox, Beatriz Rodriguez, and J David Curb. FOXO3A genotype is strongly associated with human longevity. *Proc. Natl. Acad. Sci. U. S. A.*, 105(37):13987–13992, September 2008.
- [214] Bradley J Willcox, D Craig Willcox, Qimei He, J David Curb, and Makoto Suzuki. Siblings of okinawan centenarians share lifelong mortality advantages. *J. Gerontol. A Biol. Sci. Med. Sci.*, 61(4):345–354, April 2006.

- [215] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, 46(D1):D1074–D1082, January 2018.
- [216] David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36(Database issue):D901–6, January 2008.
- [217] Mark R Woodford, Diana Dunn, Jonelle B Miller, Sami Jamal, Len Neckers, and Mehdi Mollapour. Chapter two - impact of posttranslational modifications on the anticancer activity of hsp90 inhibitors. In Jennifer Isaacs and Luke Whitesell, editors, *Advances in Cancer Research*, volume 129, pages 31–50. Academic Press, January 2016.
- [218] H Wu, C Hu, A Wang, E L Weisberg, Y Chen, C-H Yun, W Wang, Y Liu, X Liu, B Tian, J Wang, Z Zhao, Y Liang, B Li, L Wang, B Wang, C Chen, S J Buhrlage, Z Qi, F Zou, A Nonami, Y Li, S M Fernandes, S Adamia, R M Stone, I A Galinsky, X Wang, G Yang, J D Griffin, J R Brown, M J Eck, J Liu, N S Gray, and Q Liu. Discovery of a BTK/MNK dual inhibitor for lymphoma and leukemia. *Leukemia*, 30(1):173–181, January 2016.
- [219] Qinghua Wu, Wenda Wu, Vesna Jacevic, Tanos C C Franca, Xu Wang, and Kamil Kuca. Selective inhibitors for JNK signalling: a potential targeted therapy in cancer. *J. Enzyme Inhib. Med. Chem.*, 35(1):574–583, December 2020.
- [220] Aiping Xu, Zhao Zhang, Su-Hyuk Ko, Alfred L Fisher, Zhijie Liu, and Lizhen Chen. Microtubule regulators act in the nervous system to modulate fat metabolism and longevity through DAF-16 in *c. elegans*. *Aging Cell*, 18(2):e12884, April 2019.
- [221] Lingyan Xu, Xinran Ma, Narendra Verma, Luce Perie, Jay Pendse, Sama Shamloo, Anne Marie Josephson, Dongmei Wang, Jin Qiu, Mingwei Guo, Xiaodan Ping, Michele Allen, Audrey Noguchi, Danielle Springer, Fei Shen, Caizhi Liu, Shiwei Zhang, Lingyu Li, Jin Li, Junjie Xiao, Jian Lu, Zhenyu Du, Jian Luo, Jose O Aleman, Philipp Leucht, and Elisabetta Mueller. PPAR γ agonists delay age-associated metabolic disease and extend longevity. *Aging Cell*, 19(11):e13267, November 2020.
- [222] Yingxi Xu, Na Li, Rong Xiang, and Peiqing Sun. Emerging roles of the p38 MAPK and PI3K/AKT/mTOR pathways in oncogene-induced senescence. *Trends Biochem. Sci.*, 39(6):268–276, June 2014.
- [223] Xiaohui Yan, Miao Qi, Pengfei Li, Yihong Zhan, and Huanjie Shao. Apigenin in cancer therapy: anti-cancer effects and mechanisms of action. *Cell Biosci.*, 7:50, October 2017.

- [224] Ning Ye, Hengfu Yin, Jingjing Liu, Xiaogang Dai, and Tongming Yin. GESearch: An interactive GUI tool for identifying gene expression signature. *Biomed Res. Int.*, 2015:853734, June 2015.
- [225] Xiaolan Ye, James M Linton, Nicholas J Schork, Linda B Buck, and Michael Petrascheck. A pharmacological network for lifespan extension in *caenorhabditis elegans*. *Aging Cell*, 13(2):206–215, April 2014.
- [226] Jiang-An Yin, Xi-Juan Liu, Jie Yuan, Jing Jiang, and Shi-Qing Cai. Longevity manipulations differentially affect serotonin/dopamine level and behavioral deterioration in aging *caenorhabditis elegans*. *J. Neurosci.*, 34(11):3947–3958, March 2014.
- [227] Golan Yona, William Dirks, Shafquat Rahman, and David M Lin. Effective similarity measures for expression profiles. *Bioinformatics*, 22(13):1616–1622, July 2006.
- [228] Minjae Yoo, Jimin Shin, Jihye Kim, Karen A Ryall, Kyubum Lee, Sunwon Lee, Minji Jeon, Jaewoo Kang, and Aik Choon Tan. DSigDB: drug signatures database for gene set analysis. *Bioinformatics*, 31(18):3069–3071, September 2015.
- [229] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofler: an R package for comparing biological themes among gene clusters. *OMICS*, 16(5):284–287, May 2012.
- [230] Yinsong Zhu, Wenjuan He, Xiujuan Gao, Bin Li, Chenghan Mei, Rong Xu, and Hui Chen. Resveratrol overcomes gefitinib resistance by increasing the intracellular gefitinib concentration and triggering apoptosis, autophagy and senescence in PC9/G NSCLC cells. *Sci. Rep.*, 5(1):1–12, December 2015.