







ASACO: Automatic and Serial Analysis of CO-expression to discover gene modifiers with potential use in drug repurposing

Cristina Moral-Turón , Gualberto Asencio-Cortés , Francesc Rodriguez-Diaz, Alejandro Rubio, Alberto G. Navarro, Ana M. Brokate-Llanos , Andrés Garzón , Manuel J. Muñoz  and Antonio J. Pérez-Pulido 

Corresponding authors: A.J. Pérez-Pulido, Department of Molecular Biology and Biochemical Engineering (Genetics Area) University Pablo de Olavide Ctra/Utrera, Km. 1. Tel.: +34 954 348 652; E-mail: ajperez@upo.es; G. Asencio-Cortés, Department of Sport and Informatics (Computer Languages and Systems Area) University Pablo de Olavide Ctra/Utrera, Km. 1. Tel.: +34 954 967 874; E-mail: guaasecor@upo.es

Abstract

Massive gene expression analyses are widely used to find differentially expressed genes under specific conditions. The results of these experiments are often available in public databases that are undergoing a growth similar to that of molecular sequence databases in the past. This now allows novel secondary computational tools to emerge that use such information to gain new knowledge. If several genes have a similar expression profile across heterogeneous transcriptomics experiments, they could be functionally related. These associations are usually useful for the annotation of uncharacterized genes. In addition, the search for genes with opposite expression profiles is useful for finding negative regulators and proposing inhibitory compounds in drug repurposing projects. Here we present a new web application, Automatic and Serial Analysis of CO-expression (ASACO), which has the potential to discover positive and negative correlator genes to a given query gene, based on thousands of public transcriptomics experiments. In addition, examples of use are presented, comparing with previous contrasted knowledge. The results obtained propose ASACO as a useful tool to improve knowledge about genes associated with human diseases and noncoding genes. ASACO is available at <http://www.bioinfocabd.upo.es/asaco/>.

Keywords: differential gene expression; expression profile; gene annotation; noncoding genes; drug repurposing

INTRODUCTION

Differential gene expression studies enable the discovery of genes involved in particular conditions, including human diseases [1, 2]. The wide use of differential gene expression studies has led to their storage and organization in specific databases, and they now provide thousands of results that support the corresponding experiments and publications [3, 4]. These databases contain primary information and allow obtaining new knowledge, generated from the results of the stored experiments. This is the case of Expression Atlas, which uses thousands of normalized gene expression datasets to provide expression maps by tissue, and allows retrieving lists of differentially expressed genes from many diverse experiments [5].

For monogenic diseases, the results of this type of analyses allow a better understanding of how the mutations of protein-coding genes associated with the disease affect the

biological process (BP) or processes in which it participates [6]. Another type of genes for which we still have limited knowledge are noncoding RNA genes. They can regulate the amount of transcripts or proteins from protein-coding genes, and their imbalance can also cause human diseases [7, 8].

Once we know how the expression of a gene associated with a pathology changes, it may be of interest to search for drugs that restore its expression to normal physiological levels. Of particular interest in this field is drug repurposing [9, 10]. If we find out from a transcriptomics experiment that a decrease in the expression of a gene is the cause of a certain disease, and we know drugs that activate this gene expression, we could reverse the phenotype. Conversely, there are cases in which a mutation associated with a disease is not manifested in the phenotype, because a modifier gene compensates for it [11]. For example, a decrease in the expression of the modifier gene could increase the expression of

Cristina Moral-Turón was administrator of the Scientific Computing Center of the UPO (C3UPO) and current PhD student at the Institute of Biomedicine of Seville.

Gualberto Asencio-Cortés is Associate Professor at the Pablo de Olavide University.

Francesc Rodriguez-Diaz is Research Support Technician at Universidad Pablo de Olavide.

Alejandro Rubio is a near doctor from the Universidad Pablo de Olavide.

Alberto Gila Navarro was a student of the Diploma of Specialization in Bioinformatics Analysis at the Pablo de Olavide University and is currently a PhD student at the Polytechnic University of Cartagena.

Ana M. Brokate-Llanos is Adjunct Professor at the Pablo de Olavide University.

Andrés Garzón is Associate Professor at the Pablo de Olavide University and Associate Investigator at the Andalusian Center for Developmental Biology.

Manuel J. Muñoz is Associate Professor at the Pablo de Olavide University and Principal Investigator at the Andalusian Center for Developmental Biology.

Antonio J. Pérez-Pulido is Associate Professor at the Pablo de Olavide University and Principal Investigator at the Andalusian Center for Developmental Biology, head of the UPOBioinfo group.

Received: September 7, 2023. Revised: January 21, 2024. Accepted: January 31, 2024

© The Author(s) 2024. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Algorithm 1: ASACO

Input: Q: query gene; M: scoring metric; FCT: fold change threshold; PVT: p-value threshold; PN: positive and/or negative results (optional).

Output: experiments table, correlators table, enrichmentBP, enrichmentCC, enrichmentKEGG, enrichmentReactome, drugs table.

1. experiments = significantExperiments(Q, FCT, PVT, PN)
2. candidates = significantCoexpressedGenes(experiments, Q, FCT, PVT, PN)
3. correlators = scoreGenes(Q, M, candidates)
4. displayExperimentsTable(experiments)
5. displayCorrelatorsTable(correlators)
6. for each category in {"BP", "CC", "KEGG", "Reactome"}:
 - 6.1. enrichmentResults[category] = enrichment(correlators, category)
 - 6.2. displayEnrichment(category, enrichmentResults[category])
7. drugs = searchDrugs(correlators)
8. displayDrugsTable(drugs)

Algorithm 1: The ASACO algorithm implemented in the web application.

the gene associated with the disease, thus masking the phenotype. This scenario would make it possible to act on the disease by using drugs that inhibit the modifier gene.

The computational tool Automatic and Serial Analysis of CO-expression (ASACO) allows to create the expression profile of a given gene, over a series of gene expression experiments from the Expression Atlas database. ASACO has previously shown valuable results in the analysis of viral host factors [12]. Once the expression profile is created, it searches for other genes with similar expression profiles in the same experiments, called positive correlators. This can help to characterize the function of the starting gene (query gene). Alternatively, it also looks for genes with the opposite expression profile (negative correlators), which would be candidates for use as modifiers of query gene expression.

We have now created a web application that allows easy use of this bioinformatics algorithm. In this way, the user can obtain functional information about a query gene, using the biological information available (biological functions and processes) from its positive correlators. It should be noted that this information would come from various transcriptomics experiments, in which a priori the query gene was not being studied. In addition, negative correlators can be found, which, being possible negative regulators of the query gene, could act as modifiers of its expression. Likewise, by proposing a list of inhibitor drugs of these modifiers, it allows the support of future drug repurposing projects. This would allow the use of ASACO to study rare diseases where no previous transcriptomics experiments exist, proposing a list of possible modifiers and therapeutic drugs. Thus, both the disease and its associated gene could be studied based on heterogeneous independent experiments, accessible from public databases.

MATERIALS AND METHODS

ASACO algorithm

The process performed by the ASACO web application, from the moment it receives the user's query as input until all the results are produced, is explained in Algorithm 1. When we say experiment in the algorithm, we mean a comparison between two different conditions. It should be noted that sometimes we can have different experiments or comparisons within the same batch. In the latter case, we will speak of a project. Projects can include different comparisons or experimental conditions or time series, for example.

In the first step of the Algorithm 1, the function *significantExperiments* performs a search in Expression Atlas database [5] for the experiments where the query gene (Q) appears fulfilling two requirements: (a) it is expressed with a Log2-fold change equal or greater in absolute value to the specified threshold (FCT) and (b) it has a statistical significance (adjusted P-value) lower to the specified threshold (PVT). Moreover, if PN is specified, ASACO fetches only the experiment where Q has either positive or negative values of Log2-fold change. The first step returns a dictionary with the following keys: *expressionAtlasID*, *experimentID*, *experimentName*, *log2FC* and *adjustedPvalue*.

In the second step, the function *significantCoexpressedGenes* searches for all genes co-expressed differentially and significantly with Q in at least one of the experiments where this is also differentially expressed, according to the specified thresholds. If PN is provided, it considers only experiments where the query gene has either positive or negative Log2-fold change values. A dictionary is returned with the following keys: *ensembleID*, *geneName* and *log2FC*. The third step scores each candidate to

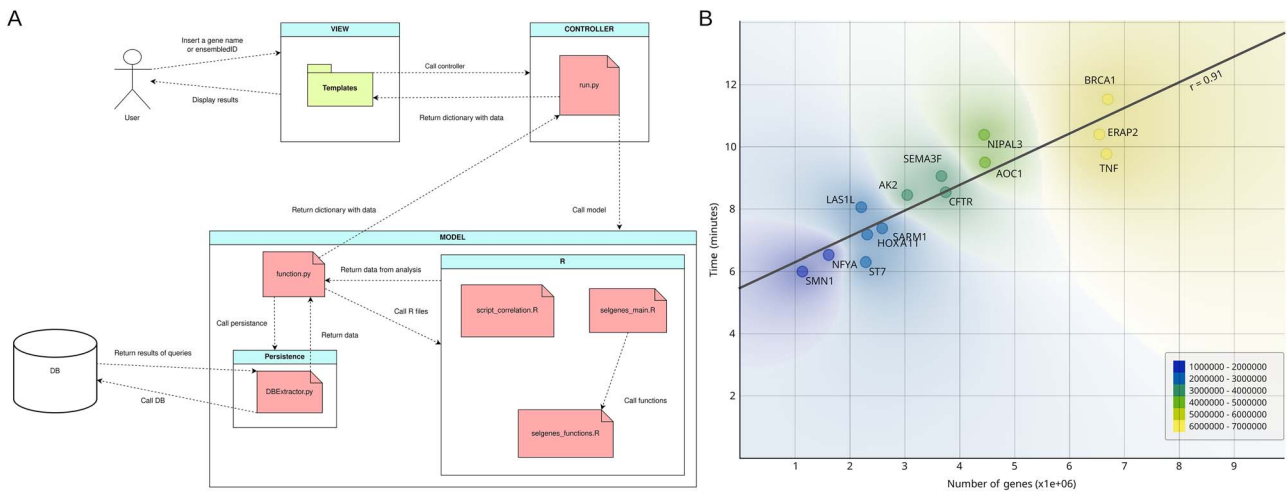


Figure 1: Implementation of the ASACO web app. (A) Component diagram of the web app. (B) Execution times (in minutes) consumed by ASACO web app for the entire process described in Algorithm 1 for a benchmark of 14 query genes. The x-axis represents the total number of genes found in all the significant experiments, before filtering out the most co-expressed candidates.

correlator according to the metric M , using Q as the reference gene. Metrics currently available in ASACO web app are: ASACO (the default metric, proposed in Reference [12]), Pearson (linear correlation coefficient), Spearman (linear correlation coefficient) and Biweight (median-based midcorrelation). A gene is proposed as a final candidate if it correlates with Q (meets the threshold of M) across all experiments considered in the first step. The fourth and fifth steps display the information of both the experiments and correlators, found in the previous steps, as tables within the web pages of results.

Step 6 performs and displays the biological enrichment of correlator genes iteratively according to each category [BP, cellular component (CC), KEGG, Reactome]. The functional enrichment of correlators are performed using GO term annotations from the BP and CC ontologies, as well as pathways from both KEGG and Reactome databases [13, 14]. The TopGO library v2.50.0 for the R language is used to perform the enrichment of GO terms, and the clusterProfiler library v4.0.2 for KEGG and Reactome pathways. Finally, steps 7 and 8 search in the DrugBank database version 5.1.5 for drugs that inhibit or activate the correlator genes and display them in a table, respectively [15].

Software architecture

ASACO web application is a client-server application implemented by using the Model-View-Controller design pattern. Specifically, Figure 1 shows the main components and their communications, indicating the most relevant information passed between them. As it is shown, the user can easily interact with the view component by specifying the query gene and receiving the final results. The view component is based on a set of web assets as templates for each web-based graphical user interface. Such a component interacts with the controller component, which is implemented in the *run.py* script. The controller interacts with the model component, which is divided in three parts: the *function.py* script, which prepares data and results to be given to the controller; the persistence subsystem, which retrieves data from the database (used in steps 1 and 2 of the ASACO algorithm); and the R subsystem, which performs the scoring candidate genes (step 3) and all the functional enrichments (step 6.1).

The application uses a SQL-based relational database (MariaDB version 10.3.32). The data model includes two entities (see

Supplementary Figure S1): experiments and genes, associated by a many-to-many relationship.

The database is stored using three tables linked by foreign keys (Supplementary Figure S2). Note that the relationship between experiments and genes is implemented by an associative table named *expressionRegisters*. Such a table contains 59 591 957 rows. Because of steps 1 and 2 of ASACO algorithm needed to retrieve data by joining the three tables, two B-tree-based indices have been created in order to improve the efficiency of the application. Specifically, the first one over the *genes_idGenes* field and the second over the *pValueAdjusted* and *log2FC* fields, both of the *expressionRegisters* table. Those indices improve the data retrieving process by 5.4 times compared with its original execution time (with no indices). The execution time in minutes consumed by ASACO web app for the entire process described in Algorithm 1 has been calculated for a benchmark of 14 query genes (Figure 1B). The results show that the linear correlation between the number of genes found and the execution time is close to 1 (specifically, $r = 0.91$). Specifically, SQL queries to the database (steps 1 and 2 of Algorithm 1) take ~8 min for the BRCA1 gene without indexes and 1.5 min with the proposed indexes. For the MYC gene, SQL queries to the database take ~98 min without indexes and 18.5 min with indexes. Note that the BRCA1 gene has 346 significant experiments found in the Expression Atlas database (step 1 of Algorithm 1), whereas the MYC gene has 523 significant experiments, requiring a more expensive execution time.

ASACO web app uses a file structure that allows it to store all the assets and scripts needed for its different functionalities (Supplementary Figure S3). The programming languages used in ASACO web app are Python, R and JavaScript. Specifically, Python has been used for main functionalities such as obtaining data from the database, calling the R files, preparing data to send to the view and creating a structure folder. R has been employed for performing biological enrichment, scoring candidate genes, and generating graphics and result files. Finally, JavaScript has been used for building download buttons, inserting data into the tables, and creating such tables along with links, among other front-end tasks.

Use cases for ASACO web app

Selection of query genes

In use case 1, we searched for a protein-coding gene with known prior functional information. For this purpose, proteins were

taken from the Swiss-Prot section of UniProtKB database [16], which had the highest score in the 'Annotation Score' field, as well as a high number of annotations from different sources: Gene Ontology terms, KEGG and Reactome pathways.

In use case 2, we searched for a noncoding gene associated with human disease and with reference functional information in the literature. Thus, the LncRNADisease v2.0 database was used, which contains 19 166 long noncoding RNA genes (lncRNA) with information on their association with human diseases [17]. The lncRNA selected was that with the highest associated relationship score with the highest possible number of human diseases, the detection method being 'Experimental'.

Results validation

To test the accuracy of the ASACO results, we take the functional annotations that the web app generates in its functional enrichment output for a gene query, including GO terms from the BP and CC ontologies, as well as KEGG and Reactome pathways. The enriched annotations obtained in this way are compared with annotations from the same previously known sources for the gene query. In cases where the ASACO annotations do not match those already known, a literature search is performed in the PubMed database, using the name of the gene query and the proposed annotation.

In use case 1 with the protein-coding gene, the correlators obtained were also compared with the interactome and other genes associated with the query gene. The BioGRID database release 4.4 [18] was used to search for proteins that were known to interact with the query protein. This database has interactomes for a total of 27,591 *Homo sapiens* proteins. Other associated genes were obtained with the 'Genes Like Me' tool of the GeneCards database version 3.12.404 [19], obtaining the first 100 genes with the highest weight based on: sequence paralogs, domains, super pathways, expression patterns, phenotypes, compounds, disorders and gene ontologies. A hypergeometric test for overrepresentation of successes in a sample (dhyper function in R) was used to calculate the P-value of the coincidence between the list of correlators of the query protein and its BioGRID interactome, as well as the match with GeneCards. To analyze the pathways associated with the matched proteins, functional enrichment was performed using the DAVID tool [20].

In use case 2 with the lncRNA gene, diseases associated with the query gene in the LncRNADisease database were also used as functional annotations for evaluation.

Finally, the results given by ASACO are accompanied by a list of correlator inhibitor drugs, extracted from the DrugBank database, which are candidates to be used for repositioning.

RESULTS

How to search for correlators with the ASACO web app

The ASACO web application allows the analysis of a query gene, given its gene name or Ensembl identifier (Figure 2A). It searches for gene expression experiments in which that gene appears as differentially expressed and assembles an expression profile (Figure 2B). It then searches for other genes with similar (positive correlator genes) and opposite (negative correlator genes) expression profiles in those same experiments.

Positive correlators are expected to be genes functionally related to the query gene, so ASACO shows an analysis of BP and pathways in which correlators are enriched (Figure 2C). In that way, the functional annotations of the query gene can

be compared with those of the correlators. In addition, the annotations of the individual correlators can be accessed by entering their record in the Ensembl database.

Finally, negative correlators could constitute modifiers of the function of the query gene. That is, the inhibition of a correlator could increase the levels of the query gene. Given this fact, ASACO also offers a list of drugs known to inhibit the correlator genes found. Thus, the user can have a set of drugs with the potential to increase the levels of the query gene.

Use case 1: example of use with a protein-coding gene linked to cancer and a rare disease

To demonstrate the utility of ASACO's web application, a protein-coding gene that was already well annotated was initially selected. Thus, the functional information of the query gene could be used as a control reference. In case the analysis worked well, the functional information of the positive correlators should match or be related to the reference.

For this gene selection, the Tumor Necrosis Factor (TNF) gene was selected (UniProt:P01375), whose protein has 136 GO terms annotated, 67 pathways in KEGG and 15 in Reactome databases (Supplementary File S1). TNF is a cell death-inducing cytokine in some tumor cell lines, mainly secreted by macrophages [21]. Under certain conditions it can also stimulate cell proliferation and induce cell differentiation. In addition, variants of this gene produce a rare disease called psoriatic arthritis [22].

Once the gene was selected, it was analyzed with ASACO. Thus, 348 experiments were found in which TNF presented a differentially gene expression (which came from 162 different transcriptomics projects), which allowed the creation of an expression profile to this gene (Figure 3A). Likewise, a total of 260 genes with an expression profile similar to TNF were found (positive correlators) and a total of 39 with an opposite expression profile (negative correlators). The expression profiles of these genes ranged between -0.6 and 0.75 , according to their Pearson correlation value (Figure 3B).

The positive correlator genes were used by ASACO to obtain a functional enrichment. Thus, the highest enrichment was found with the main pathway in which this gene is involved, TNF signaling pathway. Of the 84 genes that are part of this pathway, 32 of them were found to be positive correlators of TNF by ASACO (Figure 3C). In total, the functional enrichment found 64 KEGG pathways. Of these, 45 coincide with those already annotated for the TNF protein (Supplementary File S2). In fact, of the 67 pathways proposed by ASACO in the KEGG pathway enrichment, 45 were related to diseases or processes in which TNF is relevant (Supplementary Figure S1), with the 10 most significant among them (Figure 3D). Of particular note was the enrichment in rheumatoid arthritis, related to pathologies caused by certain TNF variants [21]. Only the term 'Cytosolic DNA-sensing pathway' was not a pathway in which TNF has been primarily implicated.

Two of the enriched pathways, not annotated for TNF protein, were 'Small cell lung cancer' and 'Acute myeloid leukemia' (Supplementary File S1). However, there is literature evidence to support the involvement of TNF in both pathways [24, 25]. This result suggests the usefulness of ASACO for proposing new pathways for a gene query.

Functional enrichment also highlights the Reactome pathway 'Signaling by interleukins', with 35 positive correlator genes, as well as GO terms related to BP in which TNF is involved (Supplementary File S1). The involvement of TNF in interleukin-mediated signaling is already known, including processes of inflammation

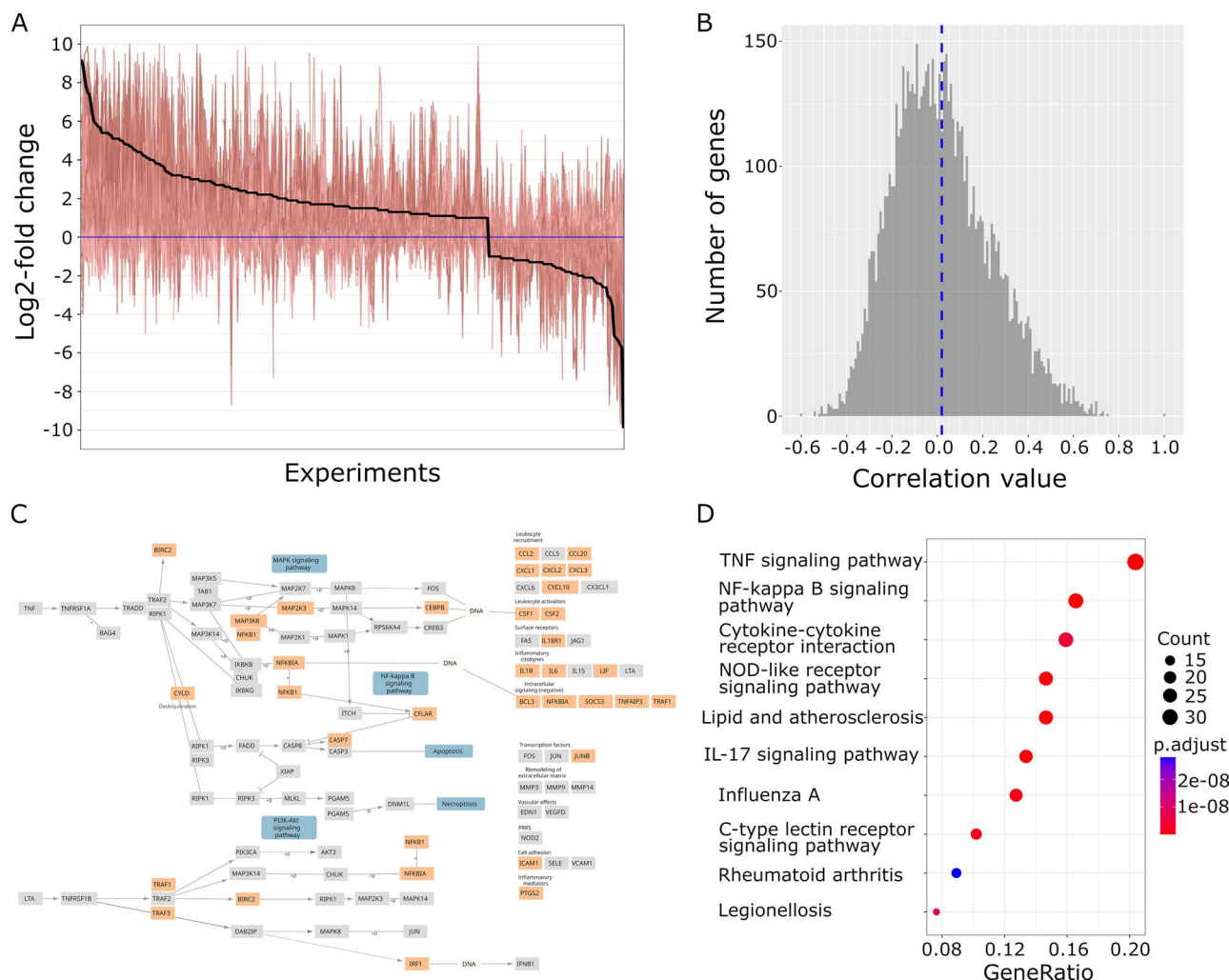


Figure 3: Results of the ASACO search for TNF-positive correlators. (A) Expression profile of 260 TNF-positive correlator genes across 348 gene expression experiments (red lines). The black line represents the TNF expression profile. (B) Pearson correlation distribution of all initial candidate genes. The blue line marks the mean correlation value. (C) TNF signaling pathway adapted from KEGG pathway, including all the genes that are part of it. Genes found by ASACO to be positive correlators are shown in orange. Relationships with other known TNF pathways are shown in blue. The pathway was drawn using the KEGGParser application of Cytoscape [23], from the KEGG KGML file; +p=phosphorylation; -p=dephosphorylation. (D) Functional enrichment of KEGG pathways for positive correlators. Note that all pathways found are associated with TNF according to KEGG Pathway database.

would be candidates to be negative modifiers of TNF levels. Therefore, known drugs against these genes could counteract these age-related effects of TNF. As an example of a TNF modifier, we have *IGF1R*, whose inhibition by interfering RNA has shown an induction in the secretion of two proinflammatory cytokines, TNF and IFNG, while delaying the growth of breast tumors in a mouse model [32]. ASACO proposes seven possible inhibitor drugs for the *IGF1R* gene, which could be useful against this type of tumors (Supplementary File S1).

Use case 2: ASACO web app also enables the discovery of correlators for noncoding genes

Noncoding genes (ncRNA) are abundant in the human genome but are still poorly understood. Knowledge of positive correlators of these genes would help to know in which BP they are involved. On the other hand, negative correlators, when present, would help to know the inhibitory functions that some of these ncRNA genes have.

To select a good candidate gene, an lncRNA gene with functional information and associations with experimental

evidence to human diseases was chosen. The gene thus chosen was *HOTAIR*, which is an lncRNA associated with different types of cancer, as well as Parkinson's disease and asthenozoospermia [33, 34]. This gene is located within the Homeobox C gene loci on chromosome 12 and is co-expressed with *HOXC* genes [35]. In addition, silencing of the *HOTAIR* gene has been associated with alterations in the organization of cellular cytoskeleton and focal adhesions [36].

Analyzing this noncoding gene with ASACO, we found 108 experiments in which it appeared differentially expressed (originating from 72 different transcriptomics projects), and 50 genes had a similar expression profile in these experiments (Figure 6A and B; Supplementary File S1). Among these positive correlators we find *HOXC9* and *HOXC11*, which are part of the *HOXC* gene cluster (Supplementary Table S1). But the main functional enrichment of these positive correlators was in intracellular transport, sperm and microtubule-based transport and movement, related to the association of *HOTAIR* with asthenozoospermia and cytoskeleton (Figure 6C and D). It also highlights a minor enrichment in regulation, which is supported

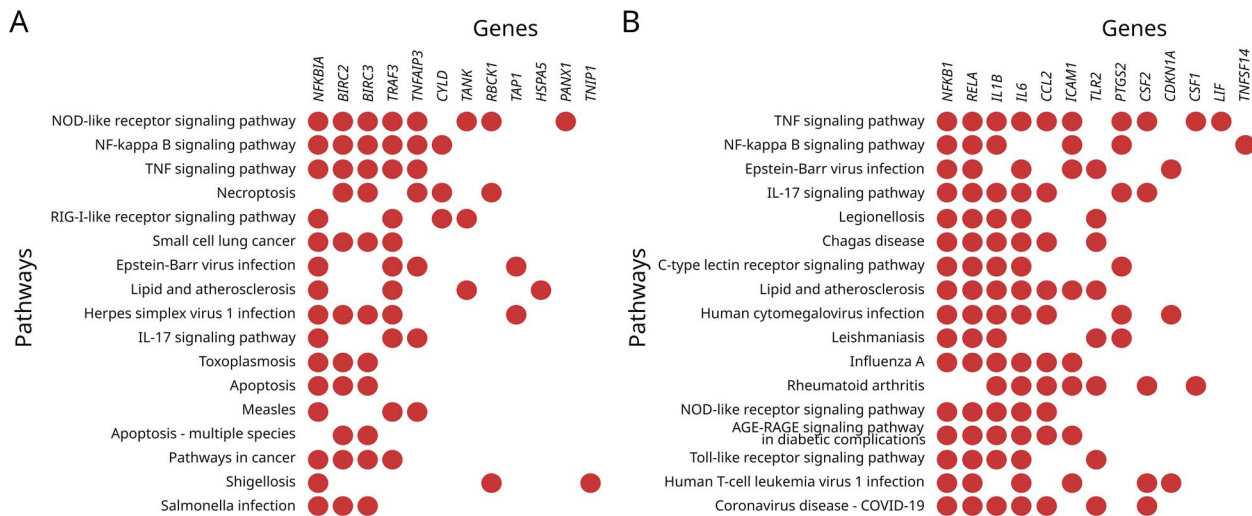


Figure 4: Pathways associated to TNF involving genes found in common by ASACO and (A) BioGRID or (B) 'Genes Like Me' from GeneCards. The pathways were found by functional enrichment using the genes found in common by both strategies. The circles represent when a gene (columns) is part of the corresponding pathway (rows). In the case of GeneCards, only the most significant pathways found are shown. See Materials and Methods for more details.

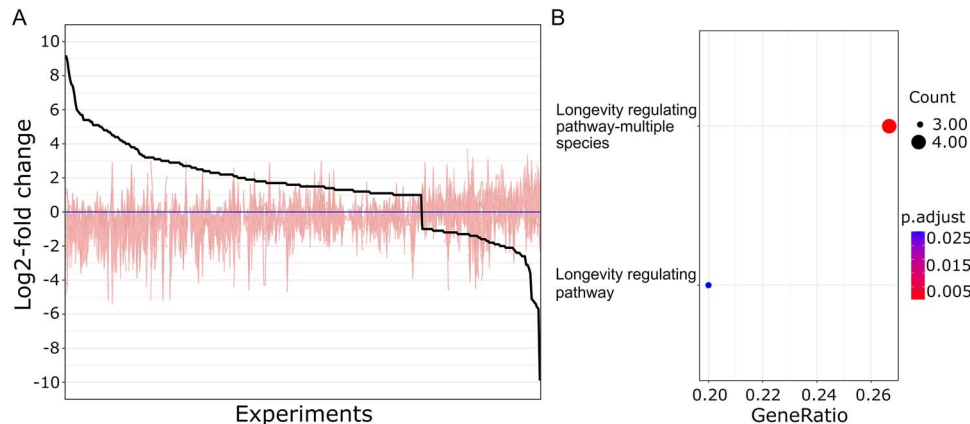


Figure 5: Results of the ASACO search for TNF-negative correlators. (A) Expression profile of 39 TNF-negative correlator genes across 728 gene expression experiments (red lines). The black line represents the TNF expression profile. (B) Functional enrichment of KEGG pathways for negative correlators.

by evidence of association of HOTAIR with mitophagy as well as Parkinson's disease [37, 38].

As for negative correlators, there are 29 genes, whose functions are particularly enriched in glycolysis, HIF-1 signaling pathway and RNA polymerase (Figure 7). HOTAIR is known for promoting glycolysis [39], and in fact, the knockdown of HOTAIR suppresses glycolysis by regulating miR-130a-3p and HIF1A in hepatocellular carcinoma cells treated by hypoxia [40]. The negative correlators found by ASACO actually point to the opposite: glycolysis would be repressed when HOTAIR levels increase. This is a contradictory result that does not agree with the current knowledge for this gene. Although the fact that it is about a pathway that has been associated with HOTAIR, it may suggest some relationship that is not currently known. However, it is important to note that this could be a false positive, and that in this case the correlators involved in this pathway should not be taken into account if we want to continue analyzing this gene further.

Finally, ASACO proposes several drugs for three of the HOTAIR negative correlators (Supplementary File S1), which added to the involvement of this gene in different types of cancer [41], could be useful for drug repurposing in this field.

DISCUSSION

Rare diseases individually affect a small part of the population. However, there are more than 10 000 known rare diseases [42], which have a high incidence when viewed as a whole. Thus, it is estimated that there are 300 million rare disease patients worldwide, which would represent about 4% of the world's population [43]. When the gene associated with a disease is known, research can be advanced, for example, by performing gene expression experiments. But this requires investment in research. In addition, things become more complicated when the target gene is poorly known, as is the case with noncoding genes [44].

Just as databases of molecular sequences long ago allowed sequence comparison to assist in homology annotation [45], now databases of gene expression experiments allow knowledge to be gained about genes that have not been specifically studied in such experiments. This enables computational analyses of understudied genes. In this context, ASACO is able to find experiments in which a specific gene has appeared differentially expressed, and the comparison of gene profiles from many of these experiments allows us to obtain functional information about that gene. This strategy can be applied to genes associated

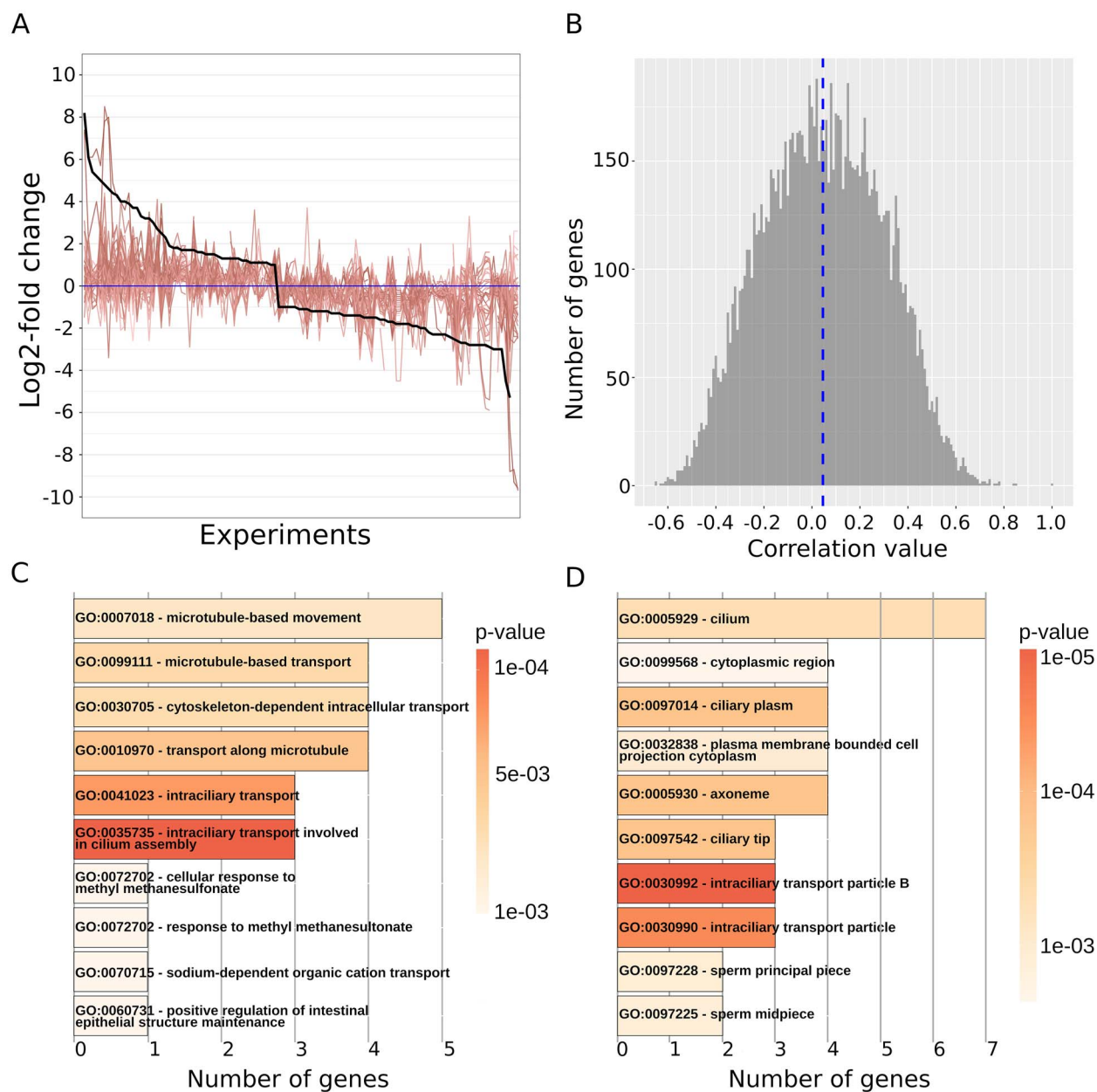


Figure 6: Results of the ASACO search for *HOTAIR*-positive correlators. (A) Expression profile of 108 *HOTAIR*-positive correlator genes across 50 gene expression experiments (red lines). The black line represents the *HOTAIR* expression profile. (B) Pearson correlation distribution of all initial candidate genes. The blue line marks the mean correlation value. Functional enrichment of the best 15 GO terms from the BP (C) and CC (D) ontology.

with diseases, or other uncharacterized genes, to extract information from transcriptomics experiments in which neither the disease nor the associated gene was being studied.

Another way to study these uncharacterized and usually understudied genes is by means of the interactome. If an uncharacterized protein is known to interact with others that are characterized, we can transfer functional information to the gene that encodes for the former. A common computational tool for this type of analysis is provided by the STRING database [46]. It allows the analysis of protein–protein interaction networks, thus helping to functionally annotate genes, which can complement information based on gene expression analysis. In fact, comparison of positive correlators of *TNF* with that of its protein interactors extracted from another similar database, BioGRID [18], gave results far superior to those expected by chance, supporting

the ASACO results. But the greatest correspondence of the positive correlators found with ASACO was with the associated genes from the GeneCards database, which are based on function, phenotype, common pathways, linked compounds, diseases and gene expression [19]. This demonstrates that ASACO can propose correlators that complement results from other sources.

ASACO is based on thousands of transcriptomics experiments, which have followed standardized processing to allow comparison between them [47]. However, these experiments are very heterogeneous and sometimes require treatment of the so-called batch effect, which leads to errors in this type of massive comparisons [48]. In the case of the genes analyzed in this work, hundreds of experiments from different sources have been considered. Expression Atlas already normalizes the experiments, and by analyzing such a large number of experiments, possible errors

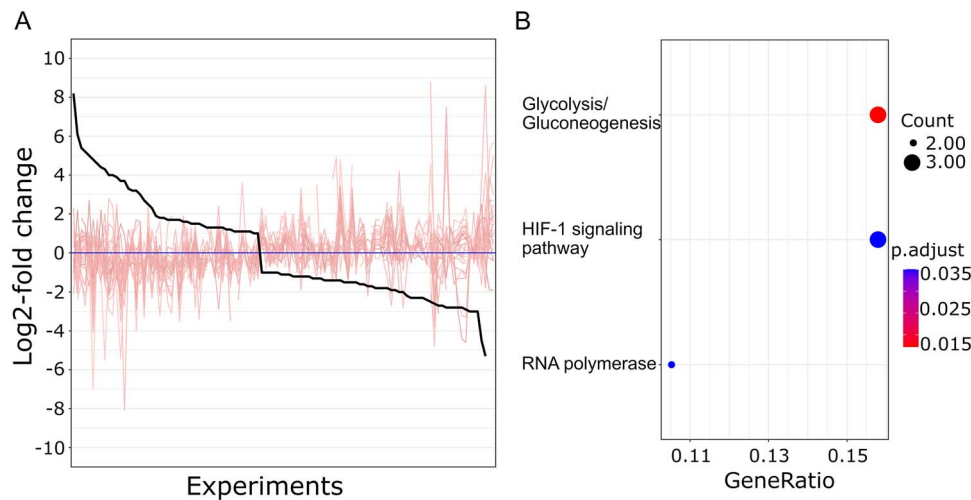


Figure 7: Results of the ASACO search for HOTAIR-negative correlators. (A) Expression profile of 29 HOTAIR-negative correlator genes across 108 gene expression experiments (red lines). The black line represents the HOTAIR expression profile. (B) Functional enrichment of KEGG pathways for negative correlators.

that may arise are smoothed out. Also, the functional enrichment of correlators carried out later helps to highlight the relevance of the results. The ASACO user can check which experiments specifically have been taken into account to search for these correlators. But in later updates of this web application, we intend to perform a better treatment of the batch effect, especially considering projects that include multiple experiments, including time series with similar conditions.

The results obtained by ASACO have been shown to be useful, because of their correspondence with the functional information previously known for these genes and contrasted with the literature. This may not only advance our knowledge of poorly studied diseases and their associated genes but would also allow functional annotation of these genes. Function assignment based on functional enrichment of gene co-expression networks is particularly useful for gaining new knowledge from massive transcriptomics data [49]. This allows the classification of experimental results as well as a better understanding of the experiment at the functional level. However, these data should be analyzed with care, and it is always recommended to review the correlator genes that have given rise to the functional enrichment, and their possible relationship with the analyzed gene.

Likewise, considering that this new knowledge is based on genes with similar or opposite expression profiles to the query gene, the use of previously known gene modifying drugs could be the basis for drug repurposing projects, not only in rare diseases, but in any human genetic disease. Thus, the selection of activator drugs for positive modifiers, or inhibitor drugs for negative modifiers, would be possible candidates for repurposing.

We believe that the post-genomic era will make an increasing number of biological experiments of all kinds available in public databases, and this will make computational tools such as ASACO, which make use of this information to obtain new knowledge, increasingly relevant.

Key Points

- Automatic and Serial Analysis of CO-expression (ASACO) allows to analyze the expression profiles of protein-coding and ncRNA genes based on thousands of public experiments.

- ASACO allows annotation of gene function from gene expression data.
- ASACO helps to find modifier genes for rare diseases.
- ASACO proposes gene expression-modifying drugs.

SUPPLEMENTARY DATA

Supplementary data are available online at <https://academic.oup.com/bfg/advance-article/doi/10.1093/bfg/ela006/7616407>

ACKNOWLEDGEMENTS

We would like to thank Naiara Landeta for her help in developing the web application.

FUNDING

No specific funding has been received for this work.

DATA AVAILABILITY

All data generated or analyzed during this study are included in this published article and its supplementary information files.

REFERENCES

- Du X-H, Ke S-B, Liang X-Y, et al. USP14 promotes colorectal cancer progression by targeting JNK for stabilization. *Cell Death Dis* 2023;**14**:56.
- Zhan L, Li J, Jew B, Sul JH. Rare variants in the endocytic pathway are associated with Alzheimer's disease, its related phenotypes, and functional consequences. *PLoS Genet* 2021;**17**:e1009772.
- Sarkans U, Gostev M, Athar A, et al. The BioStudies database—one stop shop for all data supporting a life sciences study. *Nucleic Acids Res* 2018;**46**:D1266–70.
- Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013;**41**:D991–5.
- Papatheodorou I, Moreno P, Manning J, et al. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res* 2020;**48**:D77–83.

6. Tangye SG, Gray PE, Pillay BA, et al. Hyper-IgE syndrome due to an elusive novel intronic homozygous variant in DOCK8. *J Clin Immunol* 2022;**42**:119–29.
7. Yu H, Xu Q, Liu F, et al. Identification and validation of long noncoding RNA biomarkers in human non-small-cell lung carcinomas. *J Thorac Oncol* 2015;**10**:645–54.
8. Malgundkar SH, Hassan NA, Al Badi H, et al. Identification and validation of a novel long non-coding RNA (LINC01465) in ovarian cancer. *Hum Cell* 2023;**36**:762–74.
9. Morishita EC. Discovery of RNA-targeted small molecules through the merging of experimental and computational technologies. *Expert Opin Drug Discovery* 2023;**18**:207–26.
10. Tan GSQ, Sloan EK, Lambert P, et al. Drug repurposing using real-world data. *Drug Discov Today* 2023;**28**:103422.
11. Oprea GE, Kröber S, McWhorter ML, et al. Platin 3 is a protective modifier of autosomal recessive spinal muscular atrophy. *Science* 2008;**320**:524–7.
12. Pérez-Pulido AJ, Asencio-Cortés G, Brokate-Llanos AM, et al. Serial co-expression analysis of host factors from SARS-CoV viruses highly converges with former high-throughput screenings and proposes key regulators. *Brief Bioinform* 2021;**22**:1038–52.
13. Gillespie M, Jassal B, Stephan R, et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Res* 2022;**50**:D687–92.
14. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
15. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82.
16. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**:D523–31.
17. Bao Z, Yang Z, Huang Z, et al. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res* 2019;**47**:D1034–7.
18. Oughtred R, Rust J, Chang C, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* 2021;**30**:187–200.
19. Safran M, Dalah I, Alexander J, et al. GeneCards version 3: the human gene integrator. *Database (Oxford)* 2010;**2010**:baq020.
20. Sherman BT, Hao M, Qiu J, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* 2022;**50**:W216–21.
21. Nie H, Zheng Y, Li R, et al. Phosphorylation of FOXP3 controls regulatory T cell function and is inhibited by TNF- α in rheumatoid arthritis. *Nat Med* 2013;**19**:322–8.
22. Balding J, Kane D, Livingstone W, et al. Cytokine gene polymorphisms: association with psoriatic arthritis susceptibility and severity. *Arthritis Rheum* 2003;**48**:1408–13.
23. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
24. Liu Y, Gao Y, Lin T. Expression of interleukin-1 (IL-1), IL-6, and tumor necrosis factor- α (TNF- α) in non-small cell lung cancer and its relationship with the occurrence and prognosis of cancer pain. *Ann Palliat Med* 2021;**10**:12759–66.
25. Kim MS, Kang J-W, Jeon J-S, et al. IL-32 θ gene expression in acute myeloid leukemia suppresses TNF- α production. *Oncotarget* 2015;**6**:40747–61.
26. Alam MS, Otsuka S, Wong N, et al. TNF plays a crucial role in inflammation by signaling via T cell TNFR2. *Proc Natl Acad Sci U S A* 2021;**118**:e2109972118.
27. Ishtiaq SM, Arshad MI, Khan JA. PPAR γ signaling in hepatocarcinogenesis: mechanistic insights for cellular reprogramming and therapeutic implications. *Pharmacol Ther* 2022;**240**:108298.
28. Assaraf YG. The role of multidrug resistance efflux transporters in antifolate resistance and folate homeostasis. *Drug Resist Updat* 2006;**9**:227–46.
29. Navarro-Perán E, Cabezas-Herrera J, Sánchez-del-Campo L, et al. The anti-inflammatory and anti-cancer properties of epigallocatechin-3-gallate are mediated by folate cycle disruption, adenosine release and NF- κ B suppression. *Inflamm Res* 2008;**57**:472–8.
30. Bruunsgaard H, Pedersen M, Pedersen BK. Aging and proinflammatory cytokines. *Curr Opin Hematol* 2001;**8**:131–6.
31. Bruunsgaard H, Andersen-Ranberg K, Jeune B, et al. A high plasma concentration of TNF-alpha is associated with dementia in centenarians. *J Gerontol A Biol Sci Med Sci* 1999;**54**:M357–64.
32. Durfort T, Tkach M, Meschaninova MI, et al. Small interfering RNA targeted to IGF-IR delays tumor growth and induces proinflammatory cytokines in a mouse breast cancer model. *PLoS One* 2012;**7**:e29213.
33. Sun Q, Zhang Y, Wang S, et al. LncRNA HOTAIR promotes α -synuclein aggregation and apoptosis of SH-SY5Y cells by regulating miR-221-3p in Parkinson's disease. *Exp Cell Res* 2022;**417**:113132.
34. Zhang L, Liu Z, Li X, et al. Low long non-coding RNA HOTAIR expression is associated with down-regulation of Nrf2 in the spermatozoa of patients with asthenozoospermia or oligoasthenozoospermia. *Int J Clin Exp Pathol* 2015;**8**:14198–205.
35. Rinn JL, Kertesz M, Wang JK, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 2007;**129**:1311–23.
36. Lozano-Romero A, Astudillo-de la Vega H, Terrones-Gurrola MCDR, et al. HOX transcript antisense RNA HOTAIR abrogates vasculogenic mimicry by targeting the AngiomiR-204/FAK Axis in triple negative breast cancer cells. *Noncoding RNA* 2020;**6**:19.
37. Li Y, Li W, Hoffman AR, et al. The nucleus/mitochondria-shuttling LncRNAs function as new epigenetic regulators of mitophagy in cancer. *Front Cell Dev Biol* 2021;**9**:699621.
38. Mizuno Y, Hattori N, Mori H, et al. Parkin and Parkinson's disease. *Curr Opin Neurol* 2001;**14**:477–82.
39. Wei S, Fan Q, Yang L, et al. Promotion of glycolysis by HOTAIR through GLUT1 upregulation via mTOR signaling. *Oncol Rep* 2017;**38**:1902–8.
40. Hu M, Fu Q, Jing C, et al. LncRNA HOTAIR knockdown inhibits glycolysis by regulating miR-130a-3p/HIF1A in hepatocellular carcinoma under hypoxia. *Biomed Pharmacother* 2020;**125**:109703.
41. Chen Y, Li Z, Chen X, Zhang S. Long non-coding RNAs: from disease code to drug role. *Acta Pharmaceutica Sinica B* 2021;**11**:340–54.
42. Haendel M, Vasilevsky N, Unni D, et al. How many rare diseases are there? *Nat Rev Drug Discov* 2020;**19**:77–8.
43. Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* 2020;**28**:165–73.
44. Finotti A, Fabbri E, Lampronti I, et al. MicroRNAs and long non-coding RNAs in genetic diseases. *Mol Diagn Ther* 2019;**23**:155–71.
45. Ouellette BF, Boguski MS. Database divisions and homology search files: a guide for the perplexed. *Genome Res* 1997;**7**:952–5.
46. Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein-protein association networks and functional

- enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 2023;**51**:D638–46.
47. Moreno P, Fexova S, George N, et al. Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res* 2022;**50**:D129–40.
48. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* 2017;**35**:498–507.
49. Botía JA, Vandrovcova J, Forabosco P, et al. An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. *BMC Syst Biol* 2017;**11**:47.