Computational and Structural Biotechnology Journal 21 (2023) 46–57

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Data-driven analysis and druggability assessment methods to accelerate the identification of novel cancer targets

G. Beis [a,*], A.P. Serafeim [a], I. Papasotiriou [b]

[a] Research Genetic Cancer Centre S.A., Industrial Area of Florina 53100, Greece
[b] Research Genetic Cancer Centre International GmbH Headquarters, Baarerstrasse 95, Zug 6300, Switzerland

ARTICLE INFO

Article history:
Received 26 August 2022
Received in revised form 21 November 2022
Accepted 21 November 2022
Available online 24 November 2022

Keywords:
Differential expression
Pathway analysis
Drug discovery
Druggability assessment

ABSTRACT

Over the past few decades, drug discovery has greatly improved the outcomes for patients, but several challenges continue to hinder the rapid development of novel drugs. Addressing unmet clinical needs requires the pursuit of drug targets that have a higher likelihood to lead to the development of successful drugs. Here we describe a bioinformatic approach for identifying novel cancer drug targets by performing statistical analysis to ascertain quantitative changes in expression levels between protein-coding genes, as well as co-expression networks to classify these genes into groups. Subsequently, we provide an overview of druggability assessment methodologies to prioritize and select the best targets to pursue.

Contents

1. Introduction . . . 46
2. Microarray experiments and data analysis . . . 47
 2.1. Preprocessing steps . . . 47
  2.1.1. Quality control . . . 47
  2.1.2. Probes pre-filtering . . . 48
 2.2. Normalization . . . 48
 2.3. Samples classification/clustering . . . 48
 2.4. Differential expression analysis (DEA) . . . 48
 2.5. Genes clustering and network analysis . . . 48
 2.6. Functional Annotation/ Pathway Analysis . . . 49
3. Assessing the druggability of cancer targets . . . 51
 3.1. Precedence-based assessment of druggability . . . 52
 3.2. Ligand-based assessment of druggability . . . 52
 3.3. Structure-based assessment of druggability . . . 53
 3.4. Sequence-based druggability assessment . . . 54
4. Summary and outlook . . . 54
CRediT authorship contribution statement . . . 55
Declaration of Competing Interest . . . 55
References . . . 55

* Corresponding author.
E-mail addresses: beis.giorgos@rgcc-genlab.com (G. Beis), athanasia.serafeim@rgcc-genlab.com (A.P. Serafeim), office@rgcc-genlab.com (I. Papasotiriou).

https://doi.org/10.1016/j.csbj.2022.11.042

1. Introduction

Rapid scientific advances have enabled the development of life-saving drugs, but modern-day drug discovery continues to be plagued by high costs and high attrition rates. It has been estimated that the cost for a new medicine to reach the market could be up to 1.3 billion US dollars [1]. In addition, this cost varies across different diseases, with cancer being the most expensive ailment to develop novel drugs for [2,3]. Parallel to the high cost, the success rate for drug development programs is significantly low with over 90 % of the drugs failing during clinical trials [4]. The attrition rate problem is particularly prevalent in cancer drug discovery as 95 % of the drugs that enter Phase I do not obtain marketing authorization [5].

A continuous scientific effort is put towards overcoming the challenges of drug discovery and that requires the implementation of insightful strategies throughout the development pipeline [6]. One of the critical points in developing a new drug is the preclinical stage, whose design is often not adequate for the accurate prediction of clinical efficacy and safety of a candidate drug. The reason behind this reality is that the available animal models are incomplete in mimicking human disease fully, especially human cancer [7]. Several reports have established the discrepancy between animal models and studies in humans which often results in reduced translatability of preclinical findings [8]. In cancer drug development, an intensified effort is put towards improving the in vitro and in vivo preclinical models, which can be assisted by computational cancer models, to increase confidence in their clinical relevance [9].

Another crucial point that is part of early drug discovery and needs to be addressed is target identification. Identifying disease-modifying targets and characterizing their role in the pathophysiology of the disease is the first step in the development pipeline [10]. Then, prioritizing the most promising therapeutic target minimizes the possibility of investing time and money in a poor drug target. Therefore, it is pivotal to select disease-linked and druggable targets [11]. One of the standard methods to identify novel drug targets is the microarray technology. Microarrays enable gene expression profiling and the analysis of disease versus healthy can highlight target genes of interest [12]. Having identified and validated the function of a target gene in the context of disease, another requirement is to assess the druggability of the target. Druggable targets, that can be modulated by small molecule drugs and/or biologics, are prioritized because they are more likely to result in a successful drug development project [13,14].

The strategies and decision-making during target identification can help accelerate drug development while reducing the associated costs. Identification and prioritization of the most promising therapeutic target can provide confidence in achieving a favorable outcome. In the current work, we review modern methods of processing gene expression data through statistical analysis and data mining to identify novel cancer targets. Then, we provide current methods of druggability assessment for the accurate prediction that the target can be modulated by a drug.

2. Microarray experiments and data analysis

Microarrays, as a collection of DNA probes, are generally oligonucleotides that are 'ink-jet printed' onto slides (Agilent [15]) or synthesized in situ (Affymetrix [16]). Labeled single-stranded DNA or antisense RNA fragments from a sample of interest are hybridized into the DNA microarray under high stringency conditions. The amount of hybridization detected for a specific probe is proportional to the number of nucleic acid fragments in the sample. An important design element in a microarray experiment is whether to measure the expression levels of each sample in separate microarrays (one-color arrays) or to compare the relative expression levels between a pair of samples in a single microarray (two-color arrays). The Agilent platform is primarily used to conduct the experiments, and although the Agilent panels were originally optimized for two-color analysis, a single-color protocol is now available, which includes a different panel of spike-in reagents to better optimize single-color operation [17], where each sample is labeled and hybridized to a separate microarray to obtain an absolute fluorescence value for each probe. The basic premise for microarray analysis is that relative levels of gene expression are represented by fluorescence intensities. However, the comparison cannot be made before the necessary transformations in the data have taken place to eliminate the low-intensity measurements, so that reliable comparisons can be made and differentially expressed genes can be identified with statistical significance. A fundamental microarray data design standard is governed by a respectable number of algorithms and statistical approaches that must be implemented and includes the following: quality control, probes pre-filtering, normalization, samples classification, differential expression analysis, genes clustering in combination with complex networks, and pathway analysis [18,19] (see Fig. 1).

2.1. Preprocessing steps

2.1.1. Quality control

During mRNA preprocessing, data is paramount in the path that starts from experimental design and leads to a reliable biological interpretation. Taking into account all the relevant aspects of the project, the following steps from data quality control to differential analysis leads to trustworthy results, increasing the precision and recall for prediction analysis.

Evaluation of data quality is a prerequisite before data normalization to verify whether the quality of the experimental data is acceptable for further analysis or whether any hybridization should be reconsidered [20,21]. A variety of descriptive statistics graphs is possible to identify problems with hybridization (or other experimental structures) in the quality control evaluation process.

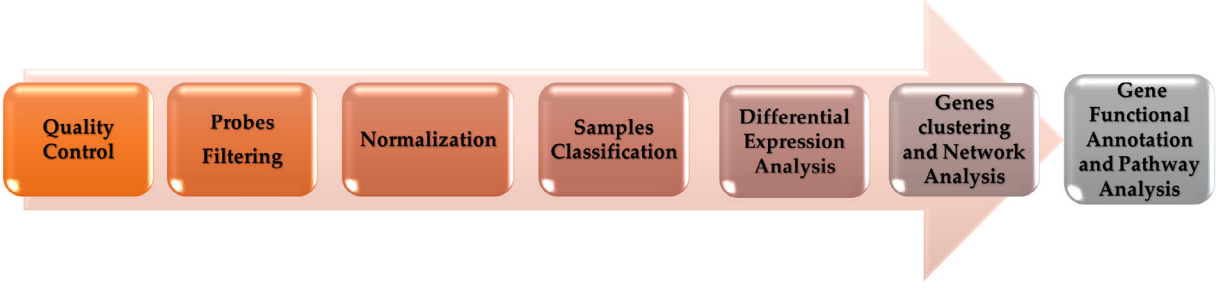

Fig. 1. Flowchart: The data analysis steps used for a microarray experiment.

Quality control charts are classified into diagnostic (i.e. MA-plots) and spot statistics (i.e. box-plots, histograms for scaling differences between different arrays or gene expressions) [21–23]. However, data preprocessing as an object, including the most common methods of visualizing, normalizing, and converting data, remains an active area of research to this day.

2.1.2. Probes pre-filtering

The accuracy of mapping microarray probes to genomic data is critical to creating reliable biological findings, but microarrays usually have a very large number of probe sets, which creates many problems in false discovery rate (FDR) control (the expected proportion of rejected null hypotheses that are false positives). This is because while biologically most sets of probes can be associated with unexpressed genes, they can be detected as differentially expressed, and therefore the measured values contain noise. Such probes should therefore be filtered and discarded from the study before further pretreatment [24].

2.2. Normalization

Microarray data tends to show high variability. Some of this variability is to be expected, as it corresponds to the differential expression of genes, but much of it arises from biases introduced during the many techniques in the experimental process. For this reason, the microarray data must first be corrected to obtain reliable intensities corresponding to the relevant level of gene expression so as to make accurate comparisons of gene expression between samples. Normalization of microarray data is the appropriate way to control the technical variation between tests while maintaining the biological variant [25]. More specifically, normalization techniques are used to reduce the variance between gene expression measurements in microarrays in order to improve data quality and the power of statistical tests for differential expression detection [26]. In essence, they are processes of scaling the raw measurement values to take into account "uninteresting" factors, for the expression levels to be more comparable both between and within the samples. Since dealing mainly with single-channel microarray data, gene expressions are normalized by "quantile normalization", which is a between-array method that seems to be most suitable for such a case [27,28]. This kind of normalization is achieved by rescaling the data distributions and fitting them to a mean distribution, in order to preserve the ranking in which the genes are ranked by expression level in each data set [29]. Another widely used preprocessing algorithm is the RMA (Robust Multi-Array Average) method [30,31] where the results are logarithmically scaled based on 2, including background correction to fluorescence intensities, normalization, as well as summary estimation of probes (since multiple probes correspond to a single gene) using Tukey's Median Polish algorithm [32].

2.3. Samples classification/clustering

Classification algorithms are generally used either to discover new classes in a data set (unsupervised classification) or to assign hypotheses to a given class (supervised classification). Here, since all the raw data have already passed the preprocessing stage, the sequence of steps includes cluster analysis with the main purpose of detecting structures in data sets. This is to discover hidden patterns in data, as well as information about which clustering technique is most suitable for distinguishing patients based on their similarity. This is important because clustering algorithms can find repeating patterns in patients that are difficult for doctors to find [33]. Cluster analyses, such as principal component analysis [34] (PCA) or hierarchical clustering [35], should be completed before further downstream analysis to ensure that samples are collected based on the experimental design. If a small number of samples show divergent clustering, their removal may be a preferred option when the sample is sufficient.

2.4. Differential expression analysis (DEA)

Subsequently, a particularly crucial step in the analysis of mRNA data is the differential expression analysis of genes (DEGs), which is a process that aims to identify those genes (that encode proteins) that are expressed at different levels between cancer states/stages and can thus provide a clearer biological picture of the processes involved in the conditions of interest. However, an additional important element for the analysis of differential gene expression, especially when comparing expressions of multiple genes and in multiple conditions for a given number of samples, is the correction of multiple tests, as in such a case we can be led to an increased probability of false-positive results, and therefore the corresponding probability values must be corrected to have a more realistic result [27]. These analyses can be achieved using computational statistical packages such as R software [36], and more specifically the Bioconductor limma package [37] which, by using ordinary linear models or ANOVA models, estimates the dependencies of covariates between samples as well as the variability of the data set. Specifically, it evaluates the differential expression, creating, among other things, an "adjusted" p-value that scores the statistical significance in order to avoid errors given by the multiple testing procedure. The classical approach to control for multiple testing is by familywise error rate (FWER), which focuses on avoiding Type I errors in a very strict way. But because it is a very conservative process for genomics, the most appropriate method of adapting multiple tests is that of Benjamin and Hochberg [38], which controls the false-discovery rate (FDR). In that way, final reporting gives in output for every gene a fold change, which explains how different the gene expression value between the conditions is, and a p-value that explains how significant is that difference. Beyond that, it is up to each researcher to set a limit on the p-value and fold-change to identify the final set of differentially expressed genes. The program also contains a variety of graphical utilities (mean-difference plots for gene expression data, etc.) for evaluating data quality, and is integrated into the Bioconductor project [39]. In addition, common reference designs are treated as single-channel for design and contrast matrices whereas empirical Bayesian methods are used to provide more consistent results. Especially, after adopting a linear model, standard errors are sufficiently mitigated using a simple Bayesian empirical model [40]. For each contrast, a moderate *t*-statistic and a log-odds probability of differential expression for each gene are also calculated.

Also, another noteworthy observation is that replicates (technical or biological) are necessary for the reliable detection of differentially expressed genes in microarray experiments. There are several methods that have been proposed to deal with the optimal number of replicates [41–43]. However, although there is no formal position on which sample-size determination procedures are best, there is an assumption that power analyses should be performed and that more replicates generally provide more power. Without repetition, statistically significant results are not possible, in the sense that there is an increased number of false positives and false-negative errors in the detection of differentially expressed genes.

2.5. Genes clustering and network analysis

A standard method of visualizing the gene expression data after the completion of the differential analysis process is to display them as a heat map based on the similarity of the gene expression pattern of the samples. This can be useful in identifying physiolog-

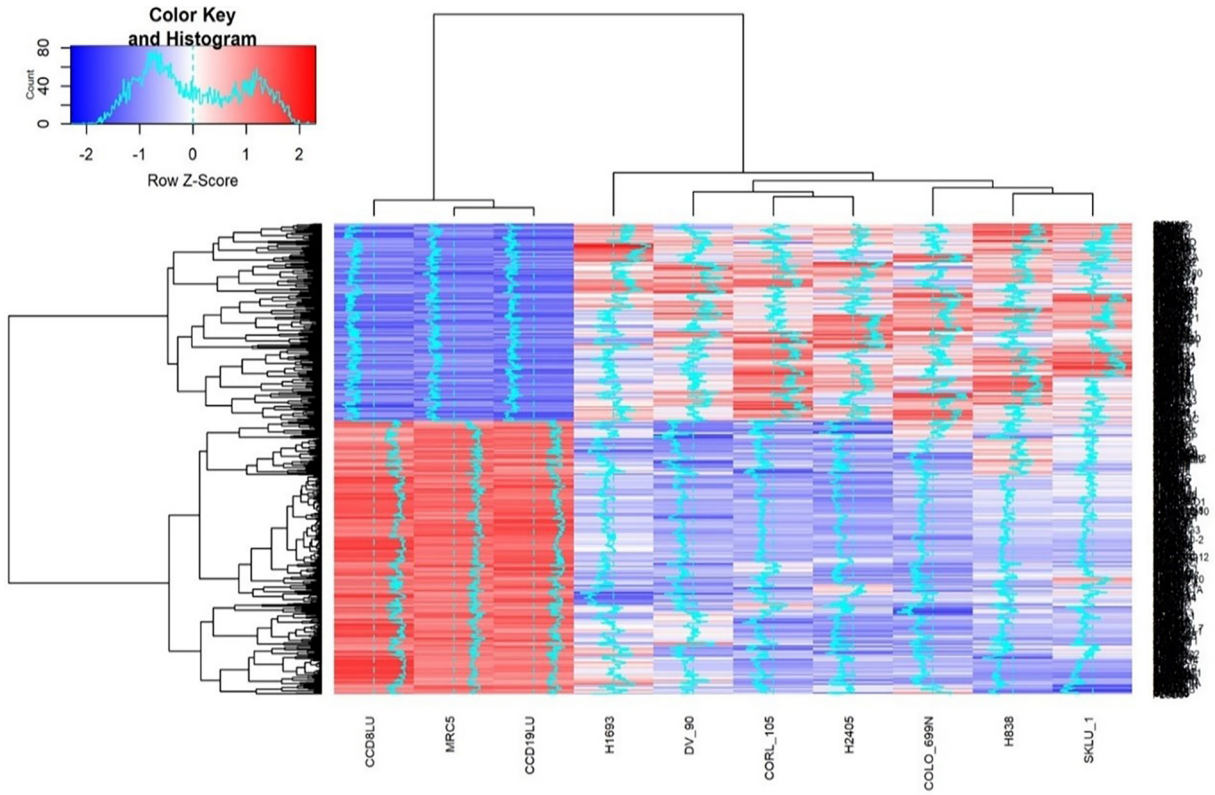

Fig. 2. A heatmap of the expression levels between treatment and controls in lung adenocarcinoma: Expression levels are shown for genes that have been identified as differentially expressed (under the conditions *|log2-fold change|>1.5* and *q-value (adjusted p-value) < 0.05*. Red and white colors indicate high and low expression levels of normalized counts by limma R package and scaled by *Z-scores* for the up-regulated genes, while blue and white colors scaled also by *Z-scores* indicate high and low expression levels for the down-regulated genes respectively. The corresponding commercial cell lines have been cultured in RGCC laboratories, from which the corresponding gene expression information was obtained and transformed on a logarithmic scale.

ically regulated genes or biological signatures associated with a particular type of cancer. In heat maps, data is displayed through hierarchical cluster analyses in a grid where each row represents a gene and each column represents a sample (see Fig. 2).

Although diseases caused by a gene variant can be detected by differential expression analysis as previously described, however, since classical microarray data analysis is based on the identification of differentially expressed genes, it is known that genes do not act alone [44]. While network analysis, in this case, can be applied to the study of gene correlation patterns in a biological system [45,46], however, the case of cancer is a complex disease caused by the variants of multiple genes and cannot be detected with the above method. In contrast, differential co-expression analysis takes into account both the interactions of multiple genes and those of specific gene pairs that are dysfunctional in cancer, as shown by a comparison of the difference in co-expression networks i.e. comparison of control vs disease [47]. Network-based co-expression studies have been used, for example, to prioritize cancer-related genes, as well as the subtype of each cancer, for the stratification of patients [48].

As differential co-expression emerged based on the analysis of gene co-expression network [49] (GCN) and assuming that related or interacting genes can share the same biological function, algorithms such as the WGCNA [45,50] that rely on the principle of 'guilty by association [51]' can be beneficial to identify co-regulated genes. By constructing co-expression networks and using normalized expression values, they can form co-expression clusters (modules). By using such a tool, topological similarity adjacency matrices (TOMs) can be constructed for a set of differentially co-expressed genes visualized as a network in order to identify genes whose expression patterns are very similar to each other and thus tend to show a coordinated expression pattern in a sample [52]. In recent years, gene co-expression networks have been mainly used to capture transcriptional patterns and predict gene interactions in functional and regulatory relationships, to provide reliable information about underlying biological functions [53]. In particular, by using these network construction tools, differentially co-expressed genes can be narrowed down into smaller gene subsets (modules) for functional term enrichment or pathway-based analyses, which will be discussed in more detail below.

2.6. Functional Annotation/ Pathway Analysis

The list of genes derived from differential expression and clustering tools are used to extrapolate biological significance from the input samples. Over-representation analysis (ORA) is a critical step of functional analysis that assesses whether a particular functionally defined gene group is not randomly represented in a set of genes that show differences in expression [54,55]. It works with Fisher's exact test [56,57], where p values represent the difference between the observed and expected overlaps of the differentially expressed genes in the experiment on the set of functional genes as well as the number of genes involved. Multiple testing correction is performed to control for false results. But as Fisher's exact test assumes that differential expression of one gene does not depend on the others (no inter-gene correlations), the test can give spurious results with long gene lists. In addition, this method is not reliable for relatively small differences.

The above problems for a more valid interpretation of the results of the transcriptomics experiments are corrected by the Gene Set Enrichment Analysis [58] approach (GSEA), which is also referred to as functional class scoring and is a rank-based threshold-free method that does not rely on differentially expressed genes to perform pathway analyses but uses all available gene expression information. The benefits of the method include the fact that it operates at the pathway level and thus, considers biological complexity, by allowing the inclusion of low-level changes that may not be detected in traditional analyses aimed at identifying differentially expressed genes. This method draws its strength by focusing on sets/groups of genes that share a common biological function. More specifically, when genetic analysis finds a little resemblance between two independent studies of patient survival in a particular type of cancer, then the GSEA reveals many common biological pathways.

Gene ontology (GO) terms, which contain standard gene annotation, are also widely used for this purpose, comparing the frequency of individual annotations in the gene list with a reference list [59,60]. Other main functional annotation databases are the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [61] that are based on Entrez gene IDs, the Reactome database of biological pathways by defining further the metabolic pathways [62], the WikiPathways [63], the Human Phenotype Ontology as a database of genes associated with human diseases [64], and the MSigDB database of gene sets [65], which in their entirety are high-performance functional genomic databases. Also, much of the data is freely available in public repositories such as ArrayExpress [66], Gene Expression Omnibus (GEO) [67], and the Comparative Toxicogenomics Database (CTD) [68]. Similarly, among PPI databases, the STRING database (https://string-db.org) provides information on the function of action and interaction at the protein level, including direct (physical) as well as indirect (functional) correlations [69]. In addition, several tools are available that represent enriched functional annotations from single pairwise comparisons, such as g: Profiler [70] and DAVID [71] where here the probability of a certain number of genes from a pathway or category in a gene list can be calculated via hypergeometric distributions [56] and Enrichr [72], but all these tools allow an analysis of only one experiment at a time. Thus, since biological processes usually involve more than one path, the result will be a network, and in the case of microarray data, is a regulatory network [73]. The role of network analysis as complementary to path analysis is to demonstrate how key elements of different pathways interact. This can be useful in identifying regulatory events that affect multiple biological processes and pathways [74].

In recent years, pathway topology [75,76] has become the first choice for extracting and explaining molecular biology for high-performance measurements, and for this reason, a large number of known knowledge path databases provide information for each path. These knowledge bases also provide information about gene products that interact with each other in a given pathway, as well as how and where they interact.

To better interpret the enrichment results, the following are examples of the most important bases for calibrating pathways and existing biological functions. More specifically, R statistical packages such as clusterProfiler [77,78], DOSE [79], ReactomePA [80], and meshes [81] are used to visualize the enrichment results accordingly (see Fig. 3).

Because dot plots show only the significantly enriched terms, and because we likely want to see exactly which genes are associated with those significant terms, it is possible to display the corresponding gene interactions and biological concepts (e.g., GO terms or KEGG pathways) in a network in order to account for the potential biological complexities in which a gene may belong to multiple annotation classes. In the following example in the results visualization, the cores below represent the corresponding biological concepts as a result of GSEA (see Fig. 4).

If the network of gene interactions becomes very complex, which can happen when there is a large number of significant terms, then a heat map is more useful in the sense that it can more easily identify expression patterns (see Fig. 5).

An additional tool in the search for gene groups with reliable biological functions is the enrichment map, which distributes

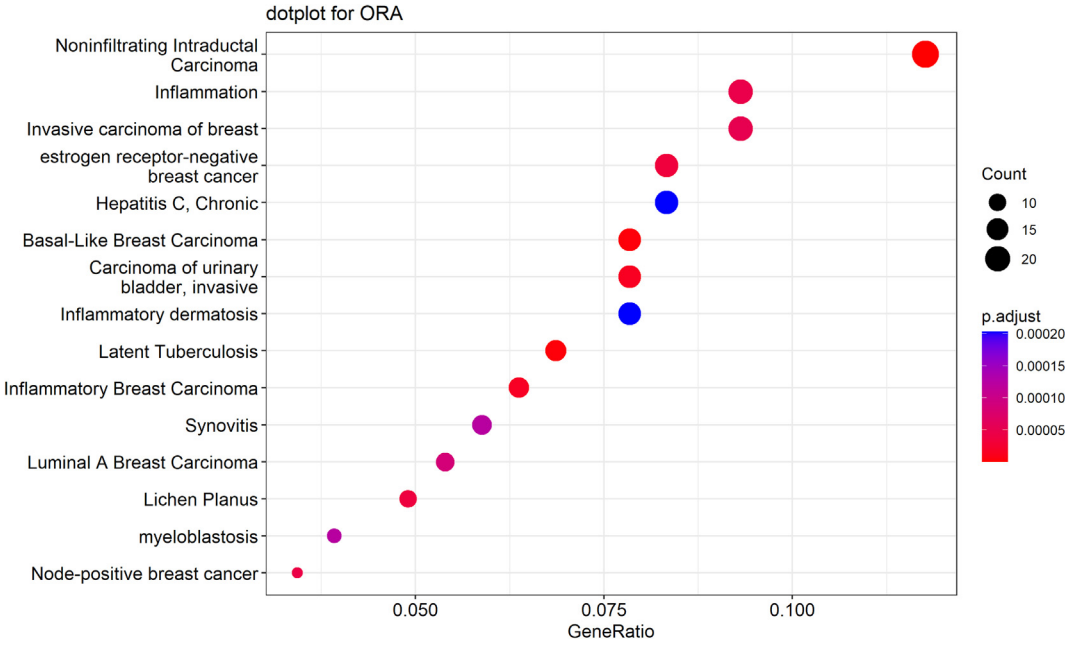

Fig. 3. The presentation of the overrepresented biological process GO terms via dot plots: The corresponding classification is derived from GeneRatio i.e. the number of condition-related genes in our selected genes divided by the corresponding number of selected genes. The p-values are adjusted by the Benjamini-Hochberg correction (BH) while the cut-off value is *0.05*. The dot size represents the number of genes in the important DE gene list associated with the GO term, while the dot color represents the adjusted p-values (BH). Visualization has been achieved via the clusterProfiler R package [77].

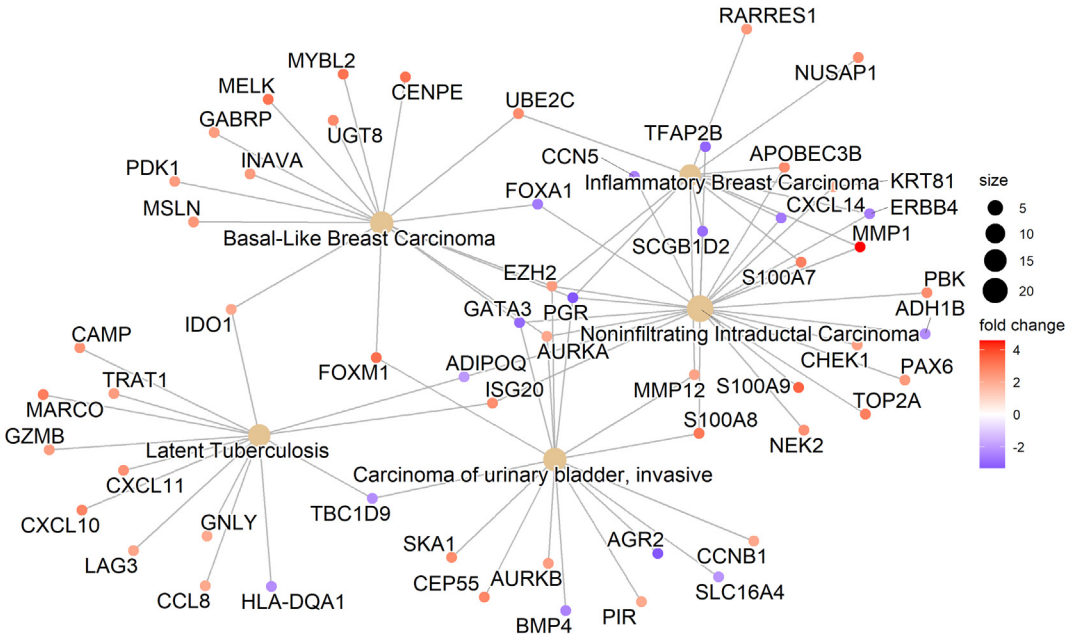

Fig. 4. The network of data relationships: The connections of genes and biological concepts (GO terms/ KEGG pathways) are depicted in the form of a network.

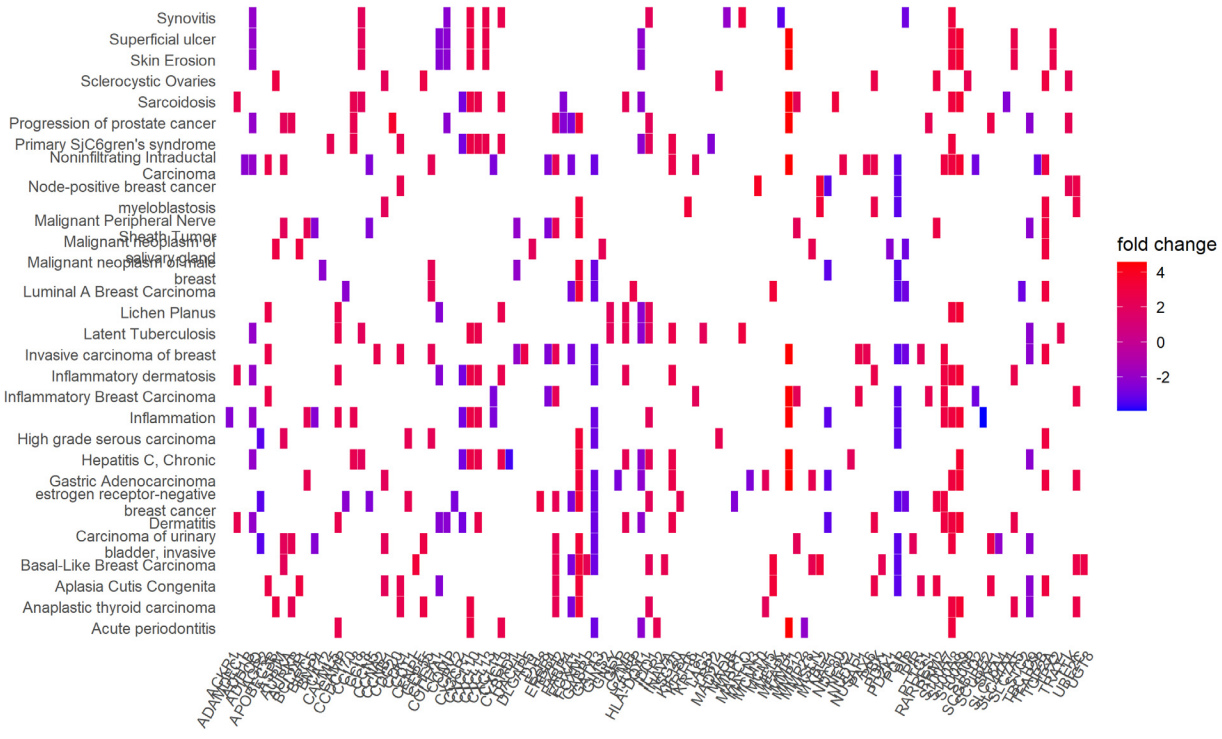

Fig. 5. The heat map for the functional classification terms used in the previous network analysis design.

enriched terms in a network that edges connect overlapping gene sets. Thus, mutually overlapping gene sets resulting in output tend to cluster together, making it easier to identify functional units. These results are obtained also via hypergeometric testing and gene set enrichment analysis (see Fig. 6).

3. Assessing the druggability of cancer targets

For a small molecule to become a marketed drug it needs to possess certain physicochemical and biological properties [82]. Therefore, the drug's target must have a binding site with complementary properties to promote the association of the two, this ability is characterized as target druggability [83]. A druggable target is most likely a protein since the great majority of FDA-approved drugs have human or pathogen-derived proteins as their target [84]. Among all the protein-coding genes in the human genome, approximately 3,000 are druggable targets and thus can be modulated by drug-like molecules [85]. Besides the experimental methods that can be deployed, various computational tools can reliably assess the druggability of novel targets. Computationally, the drug-

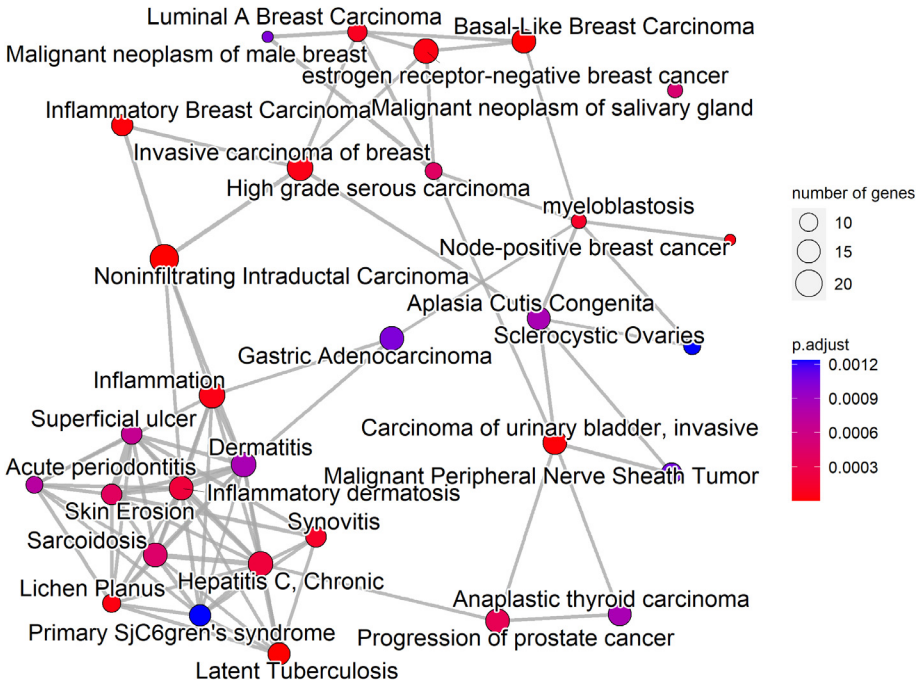

Fig. 6. The graph results from the conclusions given by the hypergeometric test and the corresponding gene set enrichment analysis.

gability can be assessed based on the target's protein sequence, 3D structure, known ligands, or any previously demonstrated druggability of the target's family members.

3.1. Precedence-based assessment of druggability

A straightforward method to predict whether a protein target is likely to be modulated by a small molecule drug is examining the druggability of its family members. A protein family consists of evolutionarily related proteins which translates into sequence, structure, and functional similarities between the members [86]. Based on these similarities, the assumption is made that when members of a protein family have demonstrated druggability, other family members may also be druggable. In the case of well-established drug targets like the major protein families of kinases, G-protein-coupled receptors (GPCRs), and ion channels, there are several reports to support the validity of a precedence-based evaluation[87–89]. For example, the multifaceted roles of GPCRs in carcinogenesis have given this protein family a prevalent role in cancer drug discovery [90]. As a result, the design of novel GPCR-based therapeutics is a plausible strategy to improve the clinical outcomes for cancer patients [91].

For a newly discovered, disease-related protein target, there are various databases that contain information on protein families and can aid with the identification of family members. Databases such as Pfam, PROSITE, CATH, and SCOP classify proteins into families based on conserved sequences, structures, and/or functions [92–95]. The collection of the target's family members is followed by searching published works and resources for any approved drugs for the family members or the target itself. Information on approved drugs and their target can be found on DrugBank and other databases [96,97]. The existence of established drug targets in the same protein family is a good indicator of druggability. If no drugs have approved that target within the family, then candidates that have reached clinical trials also provide confidence in pursuing the target. Generally, candidate drugs in late-phase trials signify that the target is safe and has an effective role in disease [98]. Information on clinical candidates can be found on ClinicalTrials.gov or the EU Clinical Trials Register [99,100].

It should be noted that the precedence-based approach for druggability assessment has limitations. The druggability of additional members of the same protein family as an existing drug target is not guaranteed. Furthermore, focusing only on traditional drug targets fails to consider protein families that have not yet been the target of drugs [101]. Expanding the so far identified druggable space could give rise to new therapies and address unmet clinical needs [102,103].

3.2. Ligand-based assessment of druggability

The ligand-based approach to assessing the druggability of a protein target applies knowledge of endogenous ligands or small molecules that bind with high affinity. Proteins naturally bind to small-molecule ligands, like adenosine triphosphate (ATP), nicotinamide adenine dinucleotide (NAD) or S-adenosyl methionine (SAM), that modulate the protein's activity [104]. The existence of such naturally occurring ligands for the target protein indicates that there is a suitable binding site to accommodate the binding of a small molecule drug [105]. Various online resources contain information on the protein's endogenous ligands such as UniProt and PubChem [106,107].

In addition to endogenous ligands, compounds designed by medicinal chemists or identified through high-throughput screens offer some confidence in the druggability of the target. Ideally, these compounds act as agonists or antagonists of their target and have demonstrated a significant binding affinity [108]. Even though they have not reached clinical trials to establish a safety and efficacy profile, primary results of drug-like small molecules can be suggestive of druggability. Information on the pharmacology of known ligands can be found on databases such as ChEMBL, BindingDB and IUPHAR/BPS Guide to PHARMACOLOGY [109–111].

A caveat of this approach is that a ligand-based assessment is not useful for novel targets for which little is known about their endogenous ligands and that have no compounds identified as binders with a promising binding affinity.

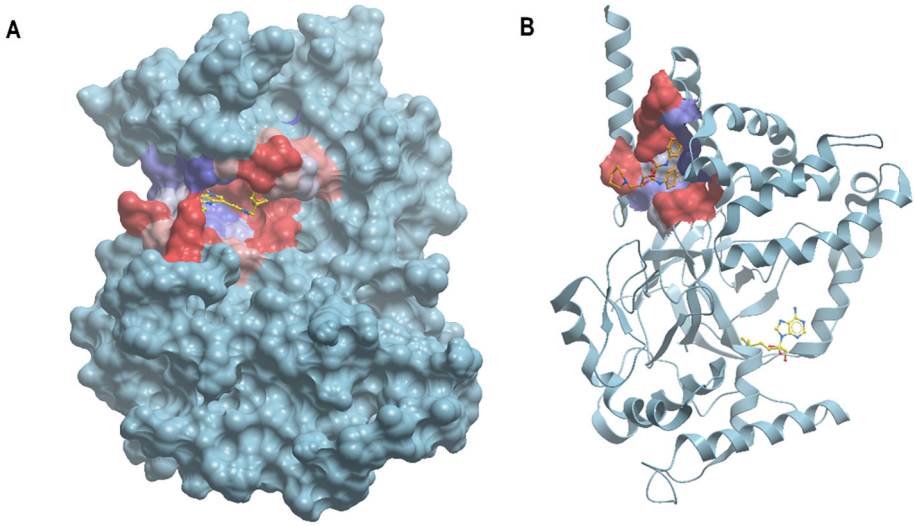

Fig. 7. 3D representations: (A) of ERK2 co-crystallized with an inhibitor (yellow) at the highlighted orthosteric pocket (PDB ID: 4qpa). (B) of SMYD3 co-crystallized with an allosteric inhibitor (orange) and a ligand bound to the orthosteric site (yellow).

3.3. Structure-based assessment of druggability

The availability of structural data for the protein target offers a great advantage in the assessment of druggability. The three-dimensional structure of a protein can be obtained from the Protein Data Bank (PDB), which presents important information regarding structural features of the target and its interaction with small molecules, peptides and other proteins [112]. The binding of a drug-like compound to a target requires the existence of a druggable pocket. Thus, the structure-based assessment inspects the protein target to detect any pockets that can be exploited for structure-based drug design [113]. The identification of a pocket initiates the search for a drug-like molecule that will block this pocket on a disease-related target therefore eliciting a therapeutic response.

'Druggable' is a pocket characterized by shape and physicochemical properties complementary to those of drug-like molecules [114]. There is no golden standard to define the most acceptable properties for a binding site to be determined as druggable. However, the general requirements are a large and deep cleft that is relatively hydrophobic in nature [115]. An example of a druggable pocket is shown in Fig. 7A, where a potent orthosteric inhibitor is bound to the active site of kinase ERK2 (PDB ID: 4qpa) [116,117]. A lot of drugs are ATP-competitive inhibitors and bind to the highly conserved ATP-binding pockets on kinases. Databases, such as HKPocket and KLIFS, are a comprehensive resource for kinase pocket structural information and kinase-ligand interactions [118,119]. Similarly, there are several online resources for the exploration of the pocketome, including PocketDB which contains predicted pockets from PDB-derived structures and CavitySpace which contains potential pockets in AlphaFold-predicted protein structures [120,121]. Additionally, a druggable pocket is not always the orthosteric site of a protein but allosteric sites have the potential to bind drug-like molecules that regulate the protein's activity [122]. A recent study by Talibov et al. identified a druggable allosteric pocket on SMYD3, a protein with implications for cancer progression (PDB ID: 6yuh, see Fig. 7B) [123,124]. The Allosteric Database contains a plethora of information that can aid with the use of allosteric druggable pockets for drug design [125]. Targeting allosteric pockets alleviates the problems of specificity that may occur when small molecules target highly conserved orthosteric sites. Finally, it should be emphasized that despite the abundance of data and resources, predicted pockets are not guaranteed to be druggable.

When there is no information on putative druggable pockets for a target of interest, various methods can be applied for the structure-based detection of suitable surface cavities on the protein. Most of the available methods are either geometry or energy-based. Geometric methods scan the protein for cavities guided by the shape and complexities of the surface whereas energy-based methods calculate the interactions of probes to identify zones of favorable binding [115]. An example of the application of several such methods, both public and commercial, is the study of Tibaut et al. which utilized them to identify pockets on the bacteriolytic enzyme autolysin E [126]. Furthermore, there are evolutionary methods that could be applied that use structure and/or sequence alignments to identify conserved regions that are likely to serve as binding pockets [115]. Examples of methods that incorporate evolutionary information include 3DLigandSite and FINDSITE [127,128].

The initial identification of binding pockets is followed by their characterization, which is a crucial step in determining which one is more likely to be druggable. Other than the search algorithm, most of the methods incorporate a scoring that is used to estimate the pocket's druggability [129]. The scoring is based on evaluating various pocket descriptors, among which the most common are volume, hydrophobic properties, and solvent accessibility [130]. Each method calculates a different combination of descriptors and the algorithm is trained on a set of known druggable targets. The resulting model can predict the druggability of the identified pockets, ideally with sufficient accuracy [131]. An example is the DLID metric used by the ICM Pocket Finder method to quantify the druggability of a protein target [132,133].

An interesting challenge that arises is that many proteins lack an adequately sized pocket in their ligand-free forms. Therefore, no pocket can be identified by the above methodologies and the target is deemed undruggable. Nevertheless, several of these proteins contain cryptic sites that only form after conformational changes induced by ligand binding [134]. Cryptic sites have been shown to bind drug-like molecules and there is significance in pursuing this type of targets for drug design [135]. Computational methods have been successful in predicting the existence of cryptic sites, with notable examples the machine learning algorithms of CryptoSite and TACTICS [135,136]. Detecting a cryptic site on a

protein target does not warrant that it is druggable, its modulation must also have an effect on the protein's function.

The structure-based assessment of druggability is limited by the need for an experimental structure of the protein target. Despite the scientific advances, only a small percentage of the proteome has an experimentally determined structure available. The machine learning algorithm of AlphaFold has contributed greatly by providing a predicted structure for 98.5 % of human proteins but only 58 % of the protein residues have a confident prediction [137]. Also considering the dynamic nature of proteins and the resulting pocket flexibility, structure-based pocket identification may not be reliable in some cases [138].

3.4. Sequence-based druggability assessment

The sequence-based approach relies on an analysis of the protein sequence to determine if the target of interest is druggable. Protein sequences can be retrieved from repositories such as RefSeq and UniProtKB [106,139]. A study by Ghadermarzi et al. identified key sequence markers that are present on drug targets and possibly druggable targets that can facilitate with the identification of novel druggable proteins [140]. These markers are sequence-derived structural and functional characteristics such as residue conservation, solvent accessibility, intrinsic disorder, alternative splicing isoforms, etc [140]. Evaluation of these markers provides a preliminary prediction of the target's druggability.

Sequences can also be utilized by machine learning algorithms to generate predictive models that can assess the druggability of targets [141]. Machine learning provides a fast prediction and most algorithms are sufficiently accurate in their performance [142]. Various sequence-derived protein features can be extracted that the algorithm learns from. It is crucial to include the most discriminatory features that can best differentiate drug targets from undruggable proteins. Furthermore, the exclusion of redundant features can help alleviate the computational cost of employing such algorithms. Recent examples of accurate predictive models include the deep learning classifier of Yu et al. and XGB-DrugPred [143,144].

4. Summary and outlook

In this review, best practices in the preprocessing of transcriptomics data derived from mRNA technologies were described. In addition, the most basic methods for performing downstream analysis were covered for single-channel microarrays that were manufactured with Agilent Technology, including normalization, differential expression analysis, and gene functional annotation. In conclusion, this review article represents a "good training" report on transcriptomics data analysis.

Potential new drug targets may include genes that are expressed differently among individuals who use anti-cancer treatment or genes that are expressed differently when a patient has been exposed to medicine known to improve or aggravate the symptoms of the disease. Also, possible targets may be genes that are co-expressed with other genes that may already be known targets and that are supposed to participate in biological systems and pathways under study. Any gene that belongs to any of these classes may be a gene for which modifying its expression can affect the cancer progression or its symptoms [145]. There are many examples of identifying such key genes that play an important role depending on each type of cancer resulting from the collection of multiple data sets and in particular from data sets of cell or tissue types most relevant to the biological target process. However, due to the limited availability of such data sets in certain regions in different cancer types, disease-specific data sets from other cell or tissue types can also be used, as they will generally not worsen the results. This is consistent with the idea of ensemble learning, in which coupling several weak and independent classifiers will lead to a strong classifier [146]. However, a more thorough evaluation of existing techniques used in bioinformatics research areas, such as classification and normalization algorithms, as well as false discovery rate estimation, is necessary.

Furthermore, the establishment of additional microarray quality control assessment methods and other new or more refined approaches to validate gene expression datasets would further benefit precise calculations for gene target discovery. To this end, reliable and well-maintained archives of datasets are required to test the validity of both older and potential new methods. These empirical data used as data for the performance and accuracy of the relevant results are based on a relatively small number of experiments, as well as a few model organisms. Proper adjustment of data preprocessing is what ensures a robust result for subsequent analysis. Therefore, special attention should be paid to this aspect in order to make a careful final decision to find and evaluate gene targets. Extensions to both other situations and species are recommended, as well as more research on how best to examine crossovers between sets of findings. Bayesian approaches [147] as a cutting-edge tool for statistical analysis could be a strong asset here, in the sense that the result will be a probability distribution rather than a point estimate. Especially when the number of replicates is small, leading to noisy point estimates, or the number of genes is very large and thus there is an extreme multiple testing problem, empirical Bayesian analysis satisfactorily corrects these classes of problems [148–150].

In addition, prioritizing and pursuing the most druggable targets contributes to successful drug discovery projects. Therefore, the assessment of druggability is a pivotal step in every drug design endeavor. Different approaches can be employed to assess the druggability of a newly identified disease-modifying protein target. Structure-based assessment yields the highest degree of confidence, but the reliability of the prediction cannot surpass the value of a 'wet lab' experiment. Thus, each prediction should be followed by experimental validation. The available algorithms and predictive tools have their strengths and limitations. The researcher benefits from the ease of use and fast results of the computational methodologies but there are caveats that need to be addressed. For example, modern algorithms and tools need to account for the intricate dynamics of drug-target interactions. Hence it is important to focus on improving the existing algorithms in terms of increasing accuracy and reducing bias. Ideally, an integration of all the available methods can boost the reliability of the prediction and the trust in pursuing the target of interest.

The current review focuses on the identification of novel cancer targets. However, the methods and resources presented can be utilized to identify targets involved in any human disorder. A typical example is the wide application of network analysis and especially WGCNA to find gene targets, such as in Alzheimer's disease or in the determination that RNF181 may be a causal gene in coronary heart disease [151,152]. A recent example where pocket prediction algorithms were utilized to identify druggable pockets is that of CRMP2. CRMP2 phosphorylation was found to be altered in various neurodegenerative diseases and is an attractive therapeutic target [153]. Another example involves the angiotensin II type 1 receptor (AT1R) for which a cryptic allosteric site that is druggable was identified through computational methods [154]. This novel allosteric pocket can be exploited to develop potent inhibitors for management of maladies like preeclampsia. These examples highlight the wide applicability of the methodologies presented in the current review.

CRediT authorship contribution statement

G. Beis: Conceptualization, Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing. **A.P. Serafeim:** Conceptualization, Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing. **I. Papasotiriou:** Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009–2018. JAMA 2020;323:844–53. https://doi.org/10.1001/jama.2020.1166.

[2] Prasad V, Mailankody S. Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval. JAMA Intern Med 2017;177:1569–75. https://doi.org/10.1001/jamainternmed.2017.3601.

[3] Leighl NB, Nirmalakumar S, Ezeife DA, Gyawali B. An Arm and a Leg: The Rising Cost of Cancer Drugs and Impact on Access. Am Soc Clin Oncol Educ Book 2021:e1–e. https://doi.org/10.1200/EDBK_100028.

[4] Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? Acta Pharm Sin B 2022. https://doi.org/10.1016/j.apsb.2022.02.002.

[5] Moreno L, Pearson ADJ. How can attrition rates be reduced in cancer drug discovery? Expert Opin Drug Discov 2013;8:363–8. https://doi.org/10.1517/17460441.2013.768984.

[6] Kiriiri GK, Njogu PM, Mwangi AN. Exploring different approaches to improve the success of drug discovery and development projects: a review. Futur J Pharm Sci 2020;6:27. https://doi.org/10.1186/s43094-020-00047-9.

[7] Kunnumakkara AB, Bordoloi D, Sailo BL, Roy NK, Thakur KK, Banik K, et al. Cancer drug development: The missing links. Exp Biol Med (Maywood) 2019;244:663–89. https://doi.org/10.1177/1535370219839163.

[8] Seyhan AA. Lost in translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. Transl Med Commun 2019;4:18. https://doi.org/10.1186/s41231-019-0050-7.

[9] Sajjad H, Imtiaz S, Noor T, Siddiqui YH, Sajjad A, Zia M. Cancer models in preclinical research: A chronicle review of advancement in effective cancer research. Animal Model Exp Med 2021;4:87–103. https://doi.org/10.1002/ame2.12165.

[10] Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol 2011;162:1239–49. https://doi.org/10.1111/j.1476-5381.2010.01127.x.

[11] Dixon SJ, Stockwell BR. Identifying druggable disease-modifying gene products. Curr Opin Chem Biol 2009;13:549–55. https://doi.org/10.1016/j.cbpa.2009.08.003.

[12] Pop L-A, Zanoaga O, Chiroi P, Nutu A, Korban SS, Stefan C, et al. Microarrays and NGS for Drug Discovery. In: Parikesit AA, editor. Drug Design. Rijeka: IntechOpen; 2021. https://doi.org/10.5772/intechopen.96657.

[13] Finan C, Gaulton A, Kruger FA, Lumbers RT, Shah T, Engmann J, et al. The druggable genome and support for target identification and validation in drug development. Sci Transl Med 2017;9:eaag1166. https://doi.org/10.1126/scitranslmed.aag1166.

[14] Dupont CA, Riegel K, Pompaiah M, Juhl H, Rajalingam K. Druggable genome and precision medicine in cancer: current challenges. FEBS J 2021;288:6142–58. https://doi.org/10.1111/febs.15788.

[15] Zhu Q, Miecznikowski JC, Halfon MS. A wholly defined Agilent microarray spike-in dataset. Bioinformatics 2011;27:1284–9. https://doi.org/10.1093/bioinformatics/btr135.

[16] Jiang N, Leach LJ, Hu X, Potokina E, Jia T, Druka A, et al. Methods for evaluating gene expression from Affymetrix microarray datasets. BMC Bioinf 2008;9:284. https://doi.org/10.1186/1471-2105-9-284.

[17] Xiao J, Lucas A, D'Andrade P, Visitacion M, Tangvoranuntakul P, Fulmer-Smentek S. Performance of the Agilent microarray platform for one-color analysis of gene expression. 2006.

[18] Federico A, Serra A, Ha MK, Kohonen P, Choi J-S, Liampa I, et al. Transcriptomics in Toxicogenomics, Part II: Preprocessing and Differential Expression Analysis for High Quality Data. Nanomaterials 2020:10. https://doi.org/10.3390/nano10050903.

[19] Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006;7:55–65. https://doi.org/10.1038/nrg1749.

[20] Barker G, Stekel D. Microarray bioinformatics. Ann Bot 2004;93:615–6. https://doi.org/10.1093/aob/mch083.

[21] Gentleman R, Carey VJ, Huber W, Irizarry RA, Doboit S. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York, NY: Springer New York; 2005. https://doi.org/10.1007/0-387-29362-0.

[22] Lee M-LT. Analysis of Microarray Gene Expression Data. Boston: Kluwer Academic Publishers; 2004. https://doi.org/10.1007/b129531.

[23] Kuyuk SA. Commonly used statistical methods for detecting differential gene expression in microarray experiments. Biostatistics and Epidemiology International Journal 2017:1–8. https://doi.org/10.30881/beji.00001.

[24] Calza S, Raffelsberger W, Ploner A, Sahel J, Leveillard T, Pawitan Y. Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. Nucleic Acids Res 2007;35:e102–e. https://doi.org/10.1093/nar/gkm537.

[25] Bolstad BM, Pre-Processing DNA, Data M. Fundamentals of Data Mining in Genomics and Proteomics. Boston, MA: Springer, US; 2007. p. 51–78.

[26] Parrish RS, Spencer HJ. Effect of Normalization on Significance Testing for Oligonucleotide Microarrays. J Biopharm Stat 2004;14:575–89. https://doi.org/10.1081/BIP-200025650.

[27] Grant GR, Manduchi E, Stoeckert CJJ. Analysis and management of microarray gene expression data. Curr Protoc Mol Biol 2007;Chapter 19:Unit 19.6. https://doi.org/10.1002/0471142727.mb1906s77.

[28] Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185–93. https://doi.org/10.1093/bioinformatics/19.2.185.

[29] Dozmorov MG, Wren JD. High-throughput processing and normalization of one-color microarrays for transcriptional meta-analyses. BMC Bioinf 2011;12:S2. https://doi.org/10.1186/1471-2105-12-S10-S2.

[30] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249–64. https://doi.org/10.1093/biostatistics/4.2.249.

[31] Irizarry RA, Wu Z, Jaffee HA. Comparison of Affymetrix GeneChip expression measures. Bioinformatics 2006;22:789–94. https://doi.org/10.1093/bioinformatics/btk046.

[32] Hoaglin DC, Mosteller F, Tukey J. Understanding Robust and Exploratory Data Analysis. 2000.

[33] Alashwal H, el Halaby M, Crouse JJ, Abdalla A, Moustafa AA. The Application of Unsupervised Clustering Methods to Alzheimer's Disease. Front Comput Neurosci 2019:13. https://doi.org/10.3389/fncom.2019.00031.

[34] Landgrebe J, Wurst W, Welzl G. Permutation-validated principal components analysis of microarray data. Genome Biol 2002;3. https://doi.org/10.1186/gb-2002-3-4-research0019.

[35] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 1998;95:14863–8. https://doi.org/10.1073/pnas.95.25.14863.

[36] Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J Comput Graph Stat 1996;5:299–314. https://doi.org/10.2307/1390807.

[37] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47–e. https://doi.org/10.1093/nar/gkv007.

[38] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc: Ser B (Methodol) 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[39] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80. https://doi.org/10.1186/gb-2004-5-10-r80.

[40] Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004;3: Article3. https://doi.org/10.2202/1544-6115.1027.

[41] Pan W, Lin J, Le CT. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biol 2002;3. https://doi.org/10.1186/gb-2002-3-5-research0022.

[42] Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics 2005;21:3017–24. https://doi.org/10.1093/bioinformatics/bti448.

[43] Müller P, Parmigiani G, Robert C, Rousseau J. Optimal Sample Size for Multiple Testing. J Am Stat Assoc 2004;99:990–1001. https://doi.org/10.1198/016214504000001646.

[44] Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinf 2010;11:95. https://doi.org/10.1186/1471-2105-11-95.

[45] Iliopoulos AC, Beis G, Apostolou P, Papasotiriou I. Complex Networks, Gene Expression and Cancer Complexity: A Brief Review of Methodology and Applications. Curr Bioinform 2020;15:629–55. https://doi.org/10.2174/1574893614666191017093504.

[46] Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert J-P. Classification of microarray data using gene networks. BMC Bioinf 2007;8:35. https://doi.org/10.1186/1471-2105-8-35.

[47] Liu B-H. Differential Coexpression Network Analysis for Gene Expression Data, 2018, p. 155–65. https://doi.org/10.1007/978-1-4939-7717-8_9.

[48] Pavel A, Serra A, Cattelani L, Federico A, Greco D. Network Analysis of Microarray Data. Methods Mol Biol 2022;2401:161–86. https://doi.org/10.1007/978-1-0716-1839-4_11.

[49] Zhang B, Horvath S. A General Framework for Weighted Gene Co-Expression Network Analysis. Stat Appl Genet Mol Biol 2005:4. https://doi.org/10.2202/1544-6115.1128.

[50] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf 2008;9:559. https://doi.org/10.1186/1471-2105-9-559.

[51] Singer GAC, Lloyd AT, Huminiecki LB, Wolfe KH. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol 2005;22:767–75. https://doi.org/10.1093/molbev/msi062.
[52] van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform 2017::bbw139. https://doi.org/10.1093/bib/bbw139.
[53] AbuQamar SF, El-Tarabily KA, Sham A. Co-expression Networks in Predicting Transcriptional Gene Regulation, 2021, p. 1–11. https://doi.org/10.1007/978-1-0716-1534-8_1.
[54] Leong HS, Kipling D. Text-based over-representation analysis of microarray gene lists with annotation bias. Nucleic Acids Res 2009;37:e79–e. https://doi.org/10.1093/nar/gkp310.
[55] García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway Analysis: State of the Art. Front Physiol 2015:6. https://doi.org/10.3389/fphys.2015.00383.
[56] Hoffman J. Hypergeometric Distribution. Biostatistics for Medical and Biomedical Practitioners, 2019, p. 734.
[57] Nguyen T-M, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol 2019;20:203. https://doi.org/10.1186/s13059-019-1790-4.
[58] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102.
[59] Dessimoz C, Škunca N, editors. The Gene Ontology Handbook. vol. 1446. New York, NY: Springer New York; 2017. https://doi.org/10.1007/978-1-4939-3743-1.
[60] Blake JA. Ten quick tips for using the gene ontology. PLoS Comput Biol 2013;9:e1003343.
[61] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27.
[62] Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 2011;39:D691–7. https://doi.org/10.1093/nar/gkq1018.
[63] Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res 2018;46:D661–7. https://doi.org/10.1093/nar/gkx1064.
[64] Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res 2017;45:D865–76. https://doi.org/10.1093/nar/gkw1039.
[65] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015;1:417–25. https://doi.org/10.1016/j.cels.2015.12.004.
[66] Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, et al. ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 2007;35:D747–50. https://doi.org/10.1093/nar/gkl995.
[67] Clough E, Barrett T. The Gene Expression Omnibus Database. Methods Mol Biol 2016;1418:93–110. https://doi.org/10.1007/978-1-4939-3578-9_5.
[68] Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, et al. Comparative Toxicogenomics Database (CTD): update 2021. Nucleic Acids Res 2021;49:D1138–43. https://doi.org/10.1093/nar/gkaa891.
[69] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47:D607–13. https://doi.org/10.1093/nar/gky1131.
[70] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019;47:W191–8. https://doi.org/10.1093/nar/gkz369.
[71] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57. https://doi.org/10.1038/nprot.2008.211.
[72] Kuleshov M. v, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Enrichr: a comprehensive gene set enrichment analysis web server, update. Nucleic Acids Res 2016;2016(44):W90–7. https://doi.org/10.1093/nar/gkw377.
[73] Werner T. Bioinformatics applications for pathway analysis of microarray data. Curr Opin Biotechnol 2008;19:50–4. https://doi.org/10.1016/j.copbio.2007.11.005.
[74] Curtis RK, Oresic M, Vidal-Puig A. Pathways to the analysis of microarray data. Trends Biotechnol 2005;23:429–35. https://doi.org/10.1016/j.tibtech.2005.05.011.
[75] Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol 2012;8:e1002375.
[76] Bayerlová M, Jung K, Kramer F, Klemm F, Bleckmann A, Beißbarth T. Comparative study on gene set and pathway topology-based enrichment methods. BMC Bioinf 2015;16:334. https://doi.org/10.1186/s12859-015-0751-5.
[77] Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7. https://doi.org/10.1089/omi.2011.0118.
[78] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation 2021;2:. https://doi.org/10.1016/j.xinn.2021.100141100141.
[79] Yu G, Wang L-G, Yan G-R, He Q-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics 2015;31:608–9. https://doi.org/10.1093/bioinformatics/btu684.
[80] Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst 2016;12:477–9. https://doi.org/10.1039/C5MB00663E.
[81] Yu G. Using meshes for MeSH term enrichment and semantic analyses. Bioinformatics 2018;34:3766–7. https://doi.org/10.1093/bioinformatics/bty410.
[82] Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 2004;1:337–41. https://doi.org/10.1016/j.ddtec.2004.11.007.
[83] Owens J. Determining druggability. Nat Rev Drug Discov 2007;6:187. https://doi.org/10.1038/nrd2275.
[84] Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov 2017;16:19–34. https://doi.org/10.1038/nrd.2016.230.
[85] Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov 2002;1:727–30. https://doi.org/10.1038/nrd892.
[86] Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families. Nat Rev Mol Cell Biol 2009;10:709–20. https://doi.org/10.1038/nrm2762.
[87] Attwood MM, Fabbro D, Sokolov A, v,, Knapp, S,, Schiöth, HB.. Trends in kinase drug discovery: targets, indications and inhibitor design. Nat Rev Drug Discov 2021;20:839–61. https://doi.org/10.1038/s41573-021-00252-y.
[88] Insel PA, Sriram K, Gorr MW, Wiley SZ, Michkov A, Salmerón C, et al. GPCRomics: An Approach to Discover GPCR Drug Targets. Trends Pharmacol Sci 2019;40:378–87. https://doi.org/10.1016/j.tips.2019.04.001.
[89] McGivern JG, Ding M. Ion Channels and Relevant Drug Screening Approaches. SLAS Discov 2020;25:413–9. https://doi.org/10.1177/2472555220921108.
[90] Chaudhary PK, Kim S. An Insight into GPCR and G-Proteins as Cancer Drivers. Cells 2021:10. https://doi.org/10.3390/cells10123288.
[91] Sriram K, Insel PA. G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? Mol Pharmacol 2018;93:251–8. https://doi.org/10.1124/mol.117.111062.
[92] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res 2021;49:D412–9. https://doi.org/10.1093/nar/gkaa913.
[93] Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Acids Res 2013;41:D344–7. https://doi.org/10.1093/nar/gks1067.
[94] Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Res 2021;49:D266–73. https://doi.org/10.1093/nar/gkaa1079.
[95] Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res 2020;48:D376–82. https://doi.org/10.1093/nar/gkz1064.
[96] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074–82. https://doi.org/10.1093/nar/gkx1037.
[97] Masoudi-Sobhanzadeh Y, Omidi Y, Amanlou M, Masoudi-Nejad A. Drug databases and their contributions to drug repurposing. Genomics 2020;112:1087–95. https://doi.org/10.1016/j.ygeno.2019.06.021.
[98] Diez PC. Clinical Drug Trials: The Path to the Patient. Methods Mol Biol 2021;2296:411–21. https://doi.org/10.1007/978-1-0716-1358-0_24.
[99] U.S. National Libray of Medicine. ClinicalTrials.gov n.d. https://www.clinicaltrials.gov/ct2/home (accessed June 20, 2022).
[100] European Medicines Agency. EU Clinical Trials Register n.d. https://www.clinicaltrialsregister.eu/ctr-search/search (accessed June 20, 2022).
[101] Patel MN, Halling-Brown MD, Tym JE, Workman P, Al-Lazikani B. Objective assessment of cancer genes for drug discovery. Nat Rev Drug Discov 2013;12:35+.
[102] Zdrazil B, Richter L, Brown N, Guha R. Moving targets in drug discovery. Sci Rep 2020;10:20213. https://doi.org/10.1038/s41598-020-77033-x.
[103] Pathmanathan S, Grozavu I, Lyakisheva A, Stagljar I. Drugging the undruggable proteins in cancer: A systems biology approach. Curr Opin Chem Biol 2022;66:. https://doi.org/10.1016/j.cbpa.2021.07.004102079.
[104] Du X, Li Y, Xia Y-L, Ai S-M, Liang J, Sang P, et al. Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. Int J Mol Sci 2016;17. https://doi.org/10.3390/ijms17020144.
[105] Bofill A, Jalencas X, Oprea TI, Mestres J. The human endogenous metabolome as a pharmacology baseline for drug discovery. Drug Discov Today 2019;24:1806–20. https://doi.org/10.1016/j.drudis.2019.06.007.
[106] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49:D480–9. https://doi.org/10.1093/nar/gkaa1100.
[107] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 2021;49:D1388–95. https://doi.org/10.1093/nar/gkaa971.
[108] Kairys V, Baranauskiene L, Kazlauskiene M, Matulis D, Kazlauskas E. Binding affinity in drug design: experimental and computational techniques. Expert

Opin Drug Discov 2019;14:755–68. https://doi.org/10.1080/17460441.2019.1623202.
[109] Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res 2017;45:D945–54. https://doi.org/10.1093/nar/gkw1074.
[110] Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 2016;44:D1045–53. https://doi.org/10.1093/nar/gkv1072.
[111] Harding SD, Armstrong JF, Faccenda E, Southan C, Alexander SPH, Davenport AP, et al. The IUPHAR/BPS guide to PHARMACOLOGY in 2022: curating pharmacology for COVID-19, malaria and antibacterials. Nucleic Acids Res 2022;50:D1282–94. https://doi.org/10.1093/nar/gkab1010.
[112] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42. https://doi.org/10.1093/nar/28.1.235.
[113] Trosset J-Y, Vodovar N. Structure-based target druggability assessment. Methods Mol Biol 2013;986:141–64. https://doi.org/10.1007/978-1-62703-311-4_10.
[114] Agoni C, Olotu FA, Ramharack P, Soliman ME. Druggability and drug-likeness concepts in drug design: are biomodelling and predictive tools having their say? J Mol Model 2020;26:120. https://doi.org/10.1007/s00894-020-04385-6.
[115] Pérot S, Sperandio O, Miteva MA, Camproux A-C, Villoutreix BO. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. Drug Discov Today 2010;15:656–67. https://doi.org/10.1016/j.drudis.2010.05.015.
[116] Burdick DJ, Wang S, Heise C, Pan B, Drummond J, Yin J, et al. Fragment-based discovery of potent ERK2 pyrrolopyrazine inhibitors. Bioorg Med Chem Lett 2015;25:4728–32. https://doi.org/10.1016/j.bmcl.2015.08.048.
[117] Burdick DJ, Wang S, Heise C, Pan B, Drummond J, Yin J, et al. Crystal Structure of ERK2 in complex with 7-(1-benzyl-1H-pyrazol-4-yl)-2-(pyridin-4-yl)-5H-pyrrolo[2,3-b]pyrazine 2014. https://doi.org/10.2210/pdb4QPA/pdb.
[118] Wang H, Qiu J, Liu H, Xu Y, Jia Y, Zhao Y. HKPocket: human kinase pocket database for drug design. BMC Bioinf 2019;20:617. https://doi.org/10.1186/s12859-019-3254-y.
[119] Kooistra AJ, Kanev GK, van Linden OPJ, Leurs R, de Esch IJP, de Graaf C. KLIFS: a structural kinase-ligand interaction database. Nucleic Acids Res 2016;44:D365–71. https://doi.org/10.1093/nar/gkv1082.
[120] Bhagavat R, Sankar S, Srinivasan N, Chandra N. An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure. Structure 2018;26:499–512.e2. https://doi.org/10.1016/j.str.2018.02.001.
[121] Wang S, Lin H, Huang Z, He Y, Deng X, Xu Y, et al. CavitySpace: A database of potential ligand binding sites in the human proteome. BioRxiv 2022. https://doi.org/10.1101/2022.01.25.477691.
[122] Song K, Zhang J. Single Binding Pockets Versus Allosteric Binding. Methods Mol Biol 2018;1825:295–326. https://doi.org/10.1007/978-1-4939-8639-2_9.
[123] Talibov VO, Fabini E, FitzGerald EA, Tedesco D, Cederfeldt D, Talu MJ, et al. Discovery of an Allosteric Ligand Binding Site in SMYD3 Lysine Methyltransferase. Chembiochem 2021;22:1597–608. https://doi.org/10.1002/cbic.202000736.
[124] Talibov VO, Fabini E, FitzGerald EA, Tedesco D, Cederfeldt D, Talu MJ, et al. Crystal structure of SMYD3 with diperodon R enantiomer bound to allosteric site 2020. https://doi.org/10.2210/pdb6YUH/pdb.
[125] Liu X, Lu S, Song K, Shen Q, Ni D, Li Q, et al. Unraveling allosteric landscapes of allosterome with ASD. Nucleic Acids Res 2020;48:D394–401. https://doi.org/10.1093/nar/gkz958.
[126] Tibaut T, Borišek J, Novič M, Turk D. Comparison of in silico tools for binding site prediction applied for structure-based design of autolysin inhibitors. SAR QSAR Environ Res 2016;27:573–87. https://doi.org/10.1080/1062936X.2016.1217271.
[127] McGreig JE, Uri H, Antczak M, Sternberg MJE, Michaelis M, Wass MN. 3DLigandSite: structure-based prediction of protein–ligand binding sites. Nucleic Acids Res 2022;50:W13–20. https://doi.org/10.1093/nar/gkac250.
[128] Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci 2008;105:129–34. https://doi.org/10.1073/pnas.0707684105.
[129] Zheng X, Gan L, Wang E, Wang J. Pocket-based drug design: exploring pocket space. AAPS J 2013;15:228–41. https://doi.org/10.1208/s12248-012-9426-6.
[130] Halgren TA. Identifying and Characterizing Binding Sites and Assessing Druggability. J Chem Inf Model 2009;49:377–89. https://doi.org/10.1021/ci800324m.
[131] Volkamer A, Kuhn D, Grombacher T, Rippmann F, Rarey M. Combining global and local measures for structure-based druggability predictions. J Chem Inf Model 2012;52:360–72. https://doi.org/10.1021/ci200454v.
[132] Sheridan RP, Maiorov VN, Holloway MK, Cornell WD, Gao Y-D. Drug-like Density: A Method of Quantifying the "Bindability" of a Protein Target Based on a Very Large Set of Pockets and Drug-like Ligands from the Protein Data Bank. J Chem Inf Model 2010;50:2029–40. https://doi.org/10.1021/ci100312t.
[133] Kufareva I, Ilatovskiy A, v,, Abagyan, R.. Pocketome: an encyclopedia of small-molecule binding sites in 4D. Nucleic Acids Res 2012;40:D535–40. https://doi.org/10.1093/nar/gkr825.
[134] Vajda S, Beglov D, Wakefield AE, Egbert M, Whitty A. Cryptic binding sites on proteins: definition, detection, and druggability. Curr Opin Chem Biol 2018;44:1–8. https://doi.org/10.1016/j.cbpa.2018.05.003.
[135] Cimermancic P, Weinkam P, Rettenmaier TJ, Bichmann L, Keedy DA, Woldeyes RA, et al. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. J Mol Biol 2016;428:709–19. https://doi.org/10.1016/j.jmb.2016.01.029.
[136] Evans DJ, Yovanno RA, Rahman S, Cao DW, Beckett MQ, Patel MH, et al. Finding Druggable Sites in Proteins using TACTICS. BioRxiv 2021:2021.02.21.432120. https://doi.org/10.1101/2021.02.21.432120.
[137] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature 2021;596:590–6. https://doi.org/10.1038/s41586-021-03828-1.
[138] Fuentes G, Dastidar SG, Madhumalar A, Verma CS. Role of protein flexibility in the discovery of new drugs. Drug Dev Res 2011;72:26–35. https://doi.org/10.1002/ddr.20399.
[139] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44:D733–45. https://doi.org/10.1093/nar/gkv1189.
[140] Ghadermarzi S, Li X, Li M, Kurgan L. Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins. Front Genet 2019;10:1075. https://doi.org/10.3389/fgene.2019.01075.
[141] Li Q, Lai L. Prediction of potential drug targets based on simple sequence properties. BMC Bioinf 2007;8:353. https://doi.org/10.1186/1471-2105-8-353.
[142] Kandoi G, Acencio ML, Lemke N. Prediction of Druggable Proteins Using Machine Learning and Systems Biology: A Mini-Review. Front Physiol 2015;6. https://doi.org/10.3389/fphys.2015.00366.
[143] Yu L, Xue L, Liu F, Li Y, Jing R, Luo J. The applications of deep learning algorithms on in silico druggable proteins identification. J Adv Res 2022. https://doi.org/10.1016/j.jare.2022.01.009.
[144] Sikander R, Ghulam A, Ali F. XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. Sci Rep 2022;12:5505. https://doi.org/10.1038/s41598-022-09484-3.
[145] Mansoori B, Mohammadi A, Davudian S, Shirjang S, Baradaran B. The Different Mechanisms of Cancer Drug Resistance: A Brief Review. Adv Pharm Bull 2017;7:339–48. https://doi.org/10.15171/apb.2017.041.
[146] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY: Springer New York; 2009. https://doi.org/10.1007/978-0-387-84858-7.
[147] Mohammad-Djafari A, Knuth KH. Bayesian approaches. Handbook of Blind Source Separation. Elsevier 2010:467–513. https://doi.org/10.1016/B978-0-12-374726-6.00017-5.
[148] Susmita D, Somnath D. Empirical Bayes screening of many p-values with applications to microarray studies. Bioinformatics 2005;21:1987–94. https://doi.org/10.1093/bioinformatics/bti301.
[149] Lewin A, Bottolo L, Richardson S. Bayesian Methods for Gene Expression Analysis. Handbook of Statistical Genomics, Wiley; 2019, p. 843–40. https://doi.org/10.1002/9781119487845.ch30.
[150] Shieh GS, Fan T-H, Chu H-P. A Bayesian approach to assessing differential expression of microarray data. J Stat Comput Simul 2008;78:179–91. https://doi.org/10.1080/10629360600954588.
[151] Liang J-W, Fang Z-Y, Huang Y, Liuyang Z, Zhang X-L, Wang J-L, et al. Application of Weighted Gene Co-Expression Network Analysis to Explore the Key Genes in Alzheimer's Disease. Journal of Alzheimer's Disease 2018;65:1353–64. https://doi.org/10.3233/JAD-180400.
[152] Dang R, Qu B, Guo K, Zhou S, Sun H, Wang W, et al. Weighted Co-Expression Network Analysis Identifies RNF181 as a Causal Gene of Coronary Artery Disease. Front Genet 2022;12. https://doi.org/10.3389/fgene.2021.818813.
[153] Khanna R, Moutal A, Perez-Miller S, Chefdeville A, Boinon L, Patek M. Druggability of CRMP2 for Neurodegenerative Diseases. ACS Chem Neurosci 2020;11:2492–505. https://doi.org/10.1021/acschemneuro.0c00307.
[154] Singh KD, Jara ZP, Harford T, Saha PP, Pardhi TR, Desnoyer R, et al. Novel allosteric ligands of the angiotensin receptor AT1R as autoantibody blockers. Proceedings of the National Academy of Sciences 2021;118. https://doi.org/10.1073/pnas.2019126118.