Computational and Structural Biotechnology Journal 20 (2022) 322–332

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

# LncRNA functional annotation with improved false discovery rate achieved by disease associations

Yongheng Wang [a,b], Jincheng Zhai [a], Xianglu Wu [c], Enoch Appiah Adu-Gyamfi [b], Lingping Yang [c], Taihang Liu [a,b], Meijiao Wang [b], Yubin Ding [b,c], Feng Zhu [d], Yingxiong Wang [b,*], Jing Tang [a,b,*]

[a] School of Basic Medicine, Chongqing Medical University, Chongqing 400016, China
[b] Joint International Research Laboratory of Reproductive and Development, Department of Reproductive Biology, School of Public Health, Chongqing Medical University, Chongqing 400016, China
[c] School of Public Health, Chongqing Medical University, Chongqing 400016, China
[d] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China

ARTICLE INFO

*Article history:*
Received 1 August 2021
Received in revised form 9 December 2021
Accepted 11 December 2021
Available online 16 December 2021

*Keywords:*
Long non-coding RNA
Functional prediction
Disease-associated SNPs
Coefficient of variation
WGCNA

ABSTRACT

The long non-coding RNAs (lncRNAs) play critical roles in various biological processes and are associated with many diseases. Functional annotation of lncRNAs in diseases attracts great attention in understanding their etiology. However, the traditional co-expression-based analysis usually produces a significant number of false positive function assignments. It is thus crucial to develop a new approach to obtain lower false discovery rate for functional annotation of lncRNAs. Here, a novel strategy termed DAnet which combining disease associations with *cis*-regulatory network between lncRNAs and neighboring protein-coding genes was developed, and the performance of DAnet was systematically compared with that of the traditional differential expression-based approach. Based on a gold standard analysis of the experimentally validated lncRNAs, the proposed strategy was found to perform better in identifying the experimentally validated lncRNAs compared with the other method. Moreover, the majority of biological pathways (40%∼100%) identified by DAnet were reported to be associated with the studied diseases. In sum, the DAnet is expected to be used to identify the function of specific lncRNAs in a particular disease or multiple diseases.



## 1. Introduction

Long non-coding RNA (lncRNA) is broadly defined as a type of non-coding RNA with a length of more than 200 nucleotides [1]. Tremendous evidences have shown that lncRNA can carry out diverse functions in biological processes [2] and is associated with many diseases [3], such as cancers [4], cardiovascular diseases [5], neurodegenerative diseases [6], metabolic diseases [7], and inflammatory diseases [8]. Currently, many computational methods for predicting lncRNA function have been developed [9], for instance, the differential expression analysis (DEA) combined with the weighted correlation network analysis (WGCNA) [10]. This method has been frequently employed for identifying co-regulatory relationships among lncRNAs and mRNAs in polycystic ovary syndrome [11] and discovering the *cis*-regulatory lncRNAs involved in vascular inflammation [12].

However, analysis based on co-expression usually results in a large number of false positive function assignments [9]. Currently, the lncRNA-disease association data supported by experiments are quite limited in the publications [13]. Specifically, only about 6,000 of over 90,000 lncRNAs have been characterized by experiments as "disease-associated" in human genome [14,15]. This may be attributed to the complex characteristics of lncRNA, including the higher expression variability across disease conditions [16–18], the susceptibility on expression/secondary structure to genetic variants [19–21], and the various levels of regulation on the coding genes (cis/trans) [2,18], etc.

So far, the analysis considering disease specificity into lncRNA functional annotation can improve the discovery of diseased-associated lncRNA [16]. In particular, lncRNA-disease associations can be well-established via the single nucleotide polymorphisms (SNPs) type of genetic variants within lncRNAs [16] and condition-specific analysis estimated by the coefficient of variation

* Corresponding authors at: School of Basic Medicine, Chongqing Medical University, Chongqing 400016, China (J. Tang).
*E-mail addresses:* yxwang@cqmu.edu.cn (Y. Wang), tang_jing@cqmu.edu.cn (J. Tang).

https://doi.org/10.1016/j.csbj.2021.12.016

**Table 1**
Twenty-four datasets of eight disease types were collected for function analysis of lncRNA. The first 22 datasets were collected from GEO and the last two datasets were collected from TCGA. MDD: major depressive disorder; VHD: valvular heart disease; AF-VHD: valvular heart disease with atrial fibrillation; SLE: systemic lupus erythematosus; ALL: acute lymphoblastic leukemia; TPM: Transcripts Per Million; Normalized: DESeq normalized; nRPKM: normalized Reads Per Kilobase of transcript, per Million mapped reads; FPKM: Fragments Per Kilobase of exon per Million; RPKM: Reads Per Kilobase of transcript per Million reads mapped; Normalized signal intensity: Quantile normalization using the GeneSpring software.

| Type of Disease | Dataset ID | No. of Sample in the specific dataset | Expression Unit (Experiment type) | No. of lncRNAs & mRNAs |
|---|---|---|---|---|
| 8A20 | GSE113524 [72] | 19 Alzheimer disease20 Healthy controls | TPM (RNA-Seq) | 12,937 lncRNAs & 18,969 mRNAs |
| 8A20 | GSE104704 [73] | 12 Alzheimer disease10 Healthy controls | Normalized (RNA-Seq) | 2,199 lncRNAs & 17,965 mRNAs |
| 8A20 | GSE125583 [74] | 219 Alzheimer disease70 Healthy controls | nRPKM (RNA-Seq) | 2,803 lncRNAs & 18,852 mRNAs |
| 6A70 | GSE101521 [75] | 30 MDD29 Healthy controls | Normalized (RNA-Seq) | 11,109 lncRNAs & 18,754 mRNAs |
| 6A70 | GSE102556 [76] | 26 MDD22 Healthy controls | FPKM (RNA-Seq) | 12,718 lncRNAs & 18,793 mRNAs |
| 6A20 | GSE112523 [77] | 29 Schizophrenia28 Healthy controls | Reads Count (RNA-Seq) | 12,179 lncRNAs & 18,437 mRNAs |
| BA41 | GSE65705 [78] | 32 Myocardial infarction2 Healthy controls | RPKM (RNA-Seq) | 1,351 lncRNAs & 17,801 mRNAs |
| BA41 | GSE127853 [79] | 3 Myocardial infarction3 Healthy controls | FPKM (RNA-Seq) | 503 lncRNAs & 10,216 mRNAs |
| BD40 | GSE97210 [80] | 3 Atherosclerosis3 Healthy controls | Normalized signal intensity (Microarray) | 10,347 lncRNAs & 18,604 mRNAs |
| BD40 | GSE120521 [81] | 4 Atherosclerosis unstable4 Atherosclerosis stable | FPKM (RNA-Seq) | 10,343 lncRNAs & 18,381 mRNAs |
| BC81 | GSE113013 [27] | 5 AF-VHD5 VHD | Normalized signal intensity (Microarray) | 10,347 lncRNAs & 18,604 mRNAs |
| BC81 | GSE108660 [27] | 5 Atrial fibrillation5 Non-atrial fibrillation | Normalized signal intensity (Microarray) | 8,090 lncRNAs & 18,807 mRNAs |
| CA23 | GSE106388 [82] | 15 Mild asthma4 Healthy controls | Reads Count (RNA-Seq) | 8,036 lncRNAs & 17,244 mRNAs |
| CA23 | GSE96783 [83] | 21 Asthma30 Healthy controls | Reads Count (RNA-Seq) | 10,451 lncRNAs & 18,324 mRNAs |
| DD71 | GSE128682 [84] | 14 Ulcerative colitis16 Healthy controls | Reads Count (RNA-Seq) | 1,756 lncRNAs & 17,355 mRNAs |
| 4A40 | GSE131525 [85] | 3 SLE3 Healthy controls | Reads Count (RNA-Seq) | 6,031 lncRNAs & 16,972 mRNAs |
| 5A10 | GSE131526 [85] | 12 Type-1 diabetes3 Healthy controls | Reads Count (RNA-Seq) | 6,798 lncRNAs & 16,458 mRNAs |
| 5B81 | GSE129398 [86] | 12 Obesity10 Controls | Reads Count (RNA-Seq) | 822 lncRNAs & 14,300 mRNAs |
| 5B81 | GSE145412 [87] | 8 Obesity8 Controls | TPM (RNA-Seq) | 6,896 lncRNAs & 16,595 mRNAs |
| 5A11 | GSE133099 [27] | 6 Type-2 diabetes6 Lean controls | Reads Count (RNA-Seq) | 8,843 lncRNAs & 17,480 mRNAs |
| 2B33 | GSE141140 [88] | 13 ALL4 Healthy controls | Reads Count (RNA-Seq) | 867 lncRNAs & 16,297 mRNAs |
| 2B91 | GSE144259 [89] | 6 Colorectal cancer3 Healthy controls | FPKM (RNA-Seq) | 3,249 lncRNAs & 18,604 mRNAs |
| 2C6Z | TCGA-BC [28] | 115 Breast cancer113 Healthy controls | FPKM (RNA-Seq) | 14,097 lncRNAs & 19,631 mRNAs |
| 2D10 | TCGA_TC [28] | 510 Thyroid cancer58 Healthy controls | Reads Count (RNA-Seq) | 13,618 lncRNAs & 19,493 mRNAs |

(CV) [17,22]. Moreover, lots of lncRNAs have been reported to regulate the expression of their neighboring genes (act in cis) [23–25]. The co-expression of the *cis*-regulatory lncRNAs and their neighboring protein-coding genes led to the discovery of functional lncRNAs in given disease [26]. It is therefore crucial to develop a new approach integrating diseased associations for obtaining lower false discovery rate (FDR) [16].

In this study, a novel strategy termed DAnet which combining disease associations with *cis*-regulatory network was developed. In particular, disease-associated SNPs were first integrated for screening disease-associated lncRNAs. And then the CV of these lncRNAs was estimated to assess the condition-specific expression of lncRNAs in a specific disease. Moreover, the WGCNA-based co-expression network between lncRNAs and their neighboring protein-coding genes and Kyoto Encyclopedia of Genes and Genome (KEGG) pathway enrichment analysis were further conducted for identifying the function of the lncRNAs involved. Furthermore, experimentally verified lncRNA-disease associations were curated to evaluate the performance of this newly proposed strategy across 24 datasets involving eight types of disease based on classification of the ICD-11. Overall, the findings of this study can facilitate the discovery of disease-associated lncRNAs and their function in the specific disease.

## 2. Methods

### 2.1. Collection of the benchmark datasets for the analysis

For the function analysis of lncRNA in different type of diseases, a variety of microarray/RNA-seq data were collected by searching disease names in Gene Expression Omnibus (GEO) [27] and The Cancer Genome Atlas (TCGA) [28]. We considered several criteria: (1) the gene expression profiling was conducted using high throughput sequencing or lncRNA microarray for "Homo sapiens", (2) the dataset consist of patient and control groups, (3) the raw data or normalized data were available, (4) the number of lncRNAs identified by disease-associated SNPs was more than zero, (5) the experimentally validated disease associated lncRNAs, which obtained from 5 public databases (LncRNAWiki [29], LncRNADisease [14], LncRNA2Target [30], Lnc2Cancer [31], and EVLncRNAs [32]), were available for the diseases and (6) multiple types of disease based on classification of the ICD-11. In total, 22 benchmark datasets were collected from GEO and two datasets were collected from TCGA, which included 16 diseases, divided into 8 types of disease according to the classification of ICD-11. Then, the lncRNA and mRNA expression matrices obtained from the 24 datasets of control-case studies were used for subsequent analysis. Table 1 demonstrates the disease type (ICD-11 code), dataset ID, the numbers of sample, the expression unit, and the number of lncRNAs and mRNAs for each dataset.

### 2.2. Collection of the SNP-disease association data for the identification of potential disease-associated lncRNAs

The SNP-disease association data were collected and used to identify potential disease-associated lncRNAs. First, we collected the 16 diseases associated SNPs and their locations from three well-known sources: GRASP2 [33], NHGRI-EBI GWAS Catalog [34], and GWASdb [35]. The significance level with $p$ less than $5.0 \times 10^{-8}$ is widely accepted in the genome-wide association studies [34]. Since many susceptible loci may only show moderate significance in association analysis, a $p$ value of less than $1.0 \times 10^{-3}$ was applied for collecting the disease-associated SNPs [35]. Then, we downloaded the chromosome information of lncRNAs from the GENCODE (v31, human reference genome hg38) [36] to map the disease-associated SNPs to the lncRNA region. In total, we collected 124,428 associations between 101,360 SNPs and the 16 diseases for further analyses, and 4,435 unique lncRNAs were found to be potentially associated with these diseases. Data details on the number of disease-associated SNPs and lncRNAs are shown in Supplementary Table S1. Finally, we exacted expression level of these lncRNAs in each dataset from raw lncRNA expression matrix, and

**Table 2**
Optimization for the $K_{CV}$ and CD across different datasets. When the $N_{exp}$ was maximum, the lower $K_{CV}$/CD was identified as the optimal value. $N_{exp}$: the number of experimental verified lncRNAs; $K_{CV:}$ the top number of lncRNAs with the higher variabilities; NA: Not available.

| Disease Name | Dataset ID | No. of lncRNA in the specific dataset | No. of lncRNA based on disease-associated SNP | No. of experimental verified lncRNA | $K_{CV}$ cutoff | CD cutoff |
|---|---|---|---|---|---|---|
| Alzheimer disease | GSE113524 | 12,937 | 1680 | 5 | 400 | 400 kb |
| Alzheimer disease | GSE104704 | 2199 | 407 | 5 | 200 | 5 kb |
| Alzheimer disease | GSE125583 | 2803 | 537 | 5 | 400 | 50 kb |
| Major depressive disorder | GSE101521 | 11,109 | 1043 | 2 | 600 | 5 kb |
| Major depressive disorder | GSE102556 | 12,718 | 1098 | 2 | 1000 | 5 kb |
| Schizophrenia | GSE112523 | 12,179 | 917 | 3 | 300 | 5 kb |
| Myocardial infarction | GSE65705 | 1351 | 35 | 2 | 35 | 100 kb |
| Myocardial infarction | GSE127853 | 503 | 16 | 2 | 16 | NA |
| Atherosclerosis | GSE97210 | 10,347 | 163 | 1 | 100 | NA |
| Atherosclerosis | GSE120521 | 10,343 | 120 | 1 | 100 | 5 kb |
| Atrial fibrillation | GSE113013 | 10,347 | 38 | 1 | 38 | NA |
| Atrial fibrillation | GSE108660 | 8090 | 33 | 1 | 33 | NA |
| Asthma | GSE106388 | 8036 | 291 | 2 | 200 | 5 kb |
| Asthma | GSE96783 | 10,451 | 352 | 2 | 100 | 5 kb |
| Lupus erythematosus | GSE131525 | 6031 | 64 | 1 | 64 | 5 kb |
| Ulcerative colitis | GSE128682 | 1756 | 20 | 1 | 20 | 70 kb |
| Type-1 diabetes mellitus | GSE131526 | 6798 | 283 | 3 | 200 | 5 kb |
| Obesity | GSE129398 | 822 | 46 | 1 | 46 | 5 kb |
| Obesity | GSE145412 | 6896 | 197 | 1 | 100 | 5 kb |
| Type-2 diabetes mellitus | GSE133099 | 8843 | 1075 | 5 | 600 | 5 kb |
| Acute lymphoblastic leukemia | GSE141140 | 867 | 12 | 1 | 12 | NA |
| Colorectal cancer | GSE144259 | 3249 | 43 | 6 | 43 | 300 kb |
| Breast cancer | TCGA_BC | 14,097 | 528 | 12 | 500 | 5 kb |
| Thyroid cancer | TCGA_TC | 13,618 | 8 | 1 | 8 | NA |

the number of the exacted lncRNAs based on disease-associated SNPs for each dataset is listed in Table 2.

### 2.3. Detection of the expression variability of lncRNA by condition-specific expression

The lncRNAs have higher expression variability pattern in diseases compared to normal conditions. LncRNAs with relative high expression variability pattern may indicate disease-related function while with relative low variability indicate function in normal condition [16,22]. The CV is the standard measurement for detecting the expression variability [16,22]. The CV is defined as "the ratio between the standard deviation of the lncRNA expression levels across the patients and its mean" [22]. In this study, we used this measurement to assess the variability of potential disease-associated lncRNAs. The CV value (ratio) was calculated for each lncRNA in disease samples, and the lncRNA with relative high CV value represents disease associated lncRNA. Finally, we ranked the CV values from high to low, and then identified the lncRNAs with top ranked CV values as the disease-associated ones. Meanwhile, different top numbers were used in the following optimization procedure. Among the top $K_{CV}$ (the top number of lncRNAs with the higher variabilities) lncRNAs across each dataset, the number of experimentally validated lncRNAs was computed ($N_{exp}$). When the number of lncRNA identified by SNPs ($N_{snp}$) was less than 100, the $K$ was equal to the $N_{snp}$, if else, the $K$ was from 100 to $N_{snp}$ with gradient of 100. When the $N_{exp}$ was maximum, the lower $K_{CV}$ was identified as the optimal value.

### 2.4. Construction of the cis-regulatory network based on lncRNAs' neighboring genes

Co-expressed genes are more likely to be co-regulated and functionally associated, meaning that identification of the co-expressed neighboring protein-coding genes can be helpful in lncRNA function assignments [16,37,38]. Firstly, we collected the information of all 16,840 lncRNAs and 19,975 protein coding genes from GENCODE (V31, human reference genome hg38) [36]. After this, we obtained 10 candidate chromosome distances (CDs) based on the publications on genomic distance between the lncRNAs and their regulated neighboring genes. These CDs including: 5 kb [39], 10 kb [40], 20 kb [41], 50 kb [42], 70 kb [43], 100 kb [44], 200 kb [45], 300 kb [46], 400 kb [47], 500 kb [12]. Secondly, we calculated the neighboring genes within these CDs up/downstream of all lncRNAs based on the collected location information. Therefore, a collection of neighboring genes of identified disease-associated lncRNAs based on SNPs and optimal $K_{CV}$ was yielded. Thirdly, we constructed the co-expression network between identified disease-associated lncRNAs and their neighboring genes in different CDs for each dataset using WGCNA [10]. Moreover, optimization procedure was performed to determine the optimal CD across the benchmark datasets. Among the lncRNAs co-expressed with neighboring genes, the number of experimentally validated lncRNAs was computed ($N_{exp}$). When the $N_{exp}$ was maximum, the lower CD was regard as the optimal one. Finally, for the functional prediction, the co-expression network based on the optimal $K_{CV}$ and CD was constructed by WGCNA for each dataset. The network of selected module identified by WGCNA was illustrated by Cytoscape 3.7.2 (http://www.cytoscape.org/) [48] software.

### 2.5. Annotating the lncRNA function based on KEGG pathway

Groups of transcripts that are identified though clustering need to be subjected to a functional enrichment step to help in revealing the biological processes that these genes are involved in [16]. The KEGG pathway [49] is globally used for characterizing the function of disease-associated lncRNA. Herein, we performed the KEGG enrichment analyses by using the mRNAs that were found to be co-expressed with disease-associated lncRNAs. The statistical significance of KEGG pathway enrichments were determined with

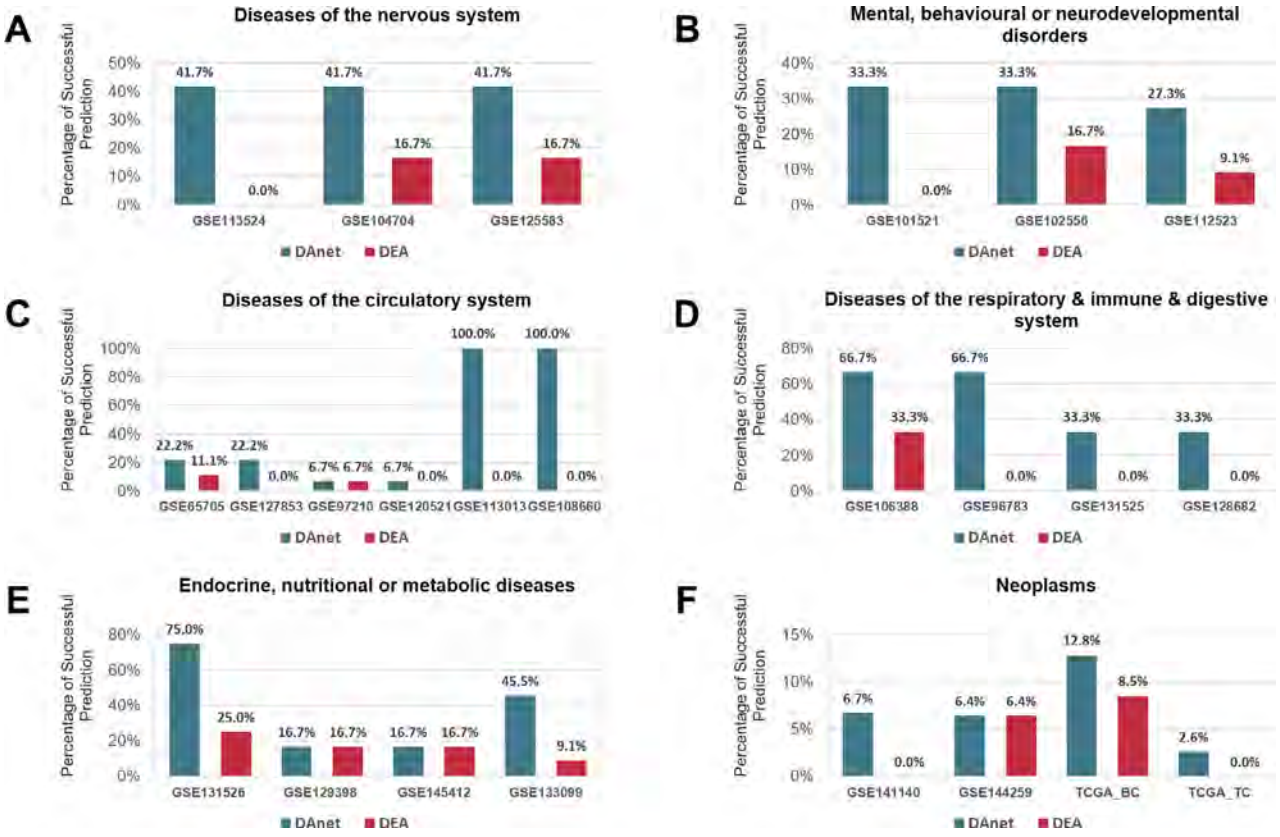


**Fig. 1.** Performance comparison between DAnet and DEA across the 24 benchmark datasets (shown in Table 1) based on the percentage of successful prediction (Rate, %), the Rate was for characterizing the experimentally verified disease associated lncRNAs.

the hypergeometric test. A $p$ value less than 0.05 indicated a significant enrichment. Also, a chord diagram was constructed using R package "circlize" [50] to illustrate the enrichment results.

### 2.6. Evaluating the ability of DAnet on the function annotation of lncRNA

As a gold standard for verifying the DAnet analysis, 9,949 pairs of experimentally verified lncRNA-disease association were integrated from five databases including LncRNAWiki [29], LncRNADisease [14], LncRNA2Target [30], Lnc2Cancer [31], and EVLncRNAs [32], which provided many experimental verified lncRNAs for diseases. Two metrics were employed to evaluate the ability of the DAnet in characterizing the function of disease-associated lncRNAs. Both metrics were based on experimentally validated disease associated lncRNAs. The metrics included: (1) percentage of successful prediction (Rate), and (2) enrichment factor (EF). The Rate (%) of DAnet and DEA (Supplementary Method S1) in characterizing the experimental verified lncRNAs was employed as the first metric to evaluate the performances. Also, EF was used to represent the comparison between the concentration of the experimentally verified lncRNAs in the identification results of DAnet/DEA and the concentration in the entire lncRNAs expression. The false discovery can be effectively evaluated by fully considering the experimentally validated disease associated lncRNAs [51]. The formula for EF is given:

$$EF = \frac{N_{truesuc}/N_{suc}}{N_{true}/N_{all}}$$

where $N_{truesuc}$ denoted the number of experimental verified lncRNAs successfully characterized as 'disease-associated' by DAnet or DEA; $N_{suc}$ represented the number of lncRNAs characterized as 'disease-associated' by DAnet or DEA; $N_{true}$ was the number of experimental verified lncRNAs in the integrated experimentally verified lncRNAs-disease associations; and $N_{all}$ indicated the total number of lncRNAs in the expression matrix. The EF no less than 1 indicated that there is an enrichment. The larger EF value represented the lower FDR [51].

## 3. Results

### 3.1. Identification of disease-specific lncRNA by SNPs across the benchmark datasets

More than 90% of disease-associated SNPs are actually located in the non-coding region (e.g., lncRNAs). The SNPs located in lncRNAs can either modify their secondary structure or affect their expression level [20]. As described in the Methods section, potential disease-associated lncRNAs of the 24 benchmark datasets were identified by disease-associated SNPs for DAnet analysis. The differential expressed lncRNAs were regarded as disease-associated lncRNAs for DEA (Supplementary Method S1). Subsequently, the Rate was utilized as a metric to measure the performance of DAnet and DEA about identifying experimentally verified lncRNAs. As shown in Supplementary Fig. S1, the Rate value of each dataset by the adjusted p value (from 0% for 18 datasets to 16.7% for GSE125583) was lower than that by the $p$ value (from 0% for 11 datasets to 33.3% for GSE106388). Among the 24 datasets, there were 8 datasets with no differentially expressed genes using the FDR less than 0.05. Thus, the raw $p$ value ($p$ less than 0.05) was used for identifying the differential expressed lncRNAs across the 24 datasets.

As shown in Fig. 1, the Rate of DAnet was varied (from 2.6% for TCGA-TC to 100% for GSE113013 and GSE108660) and the Rate of DEA was also differed greatly (from 0% for 11 datasets to 33.3%

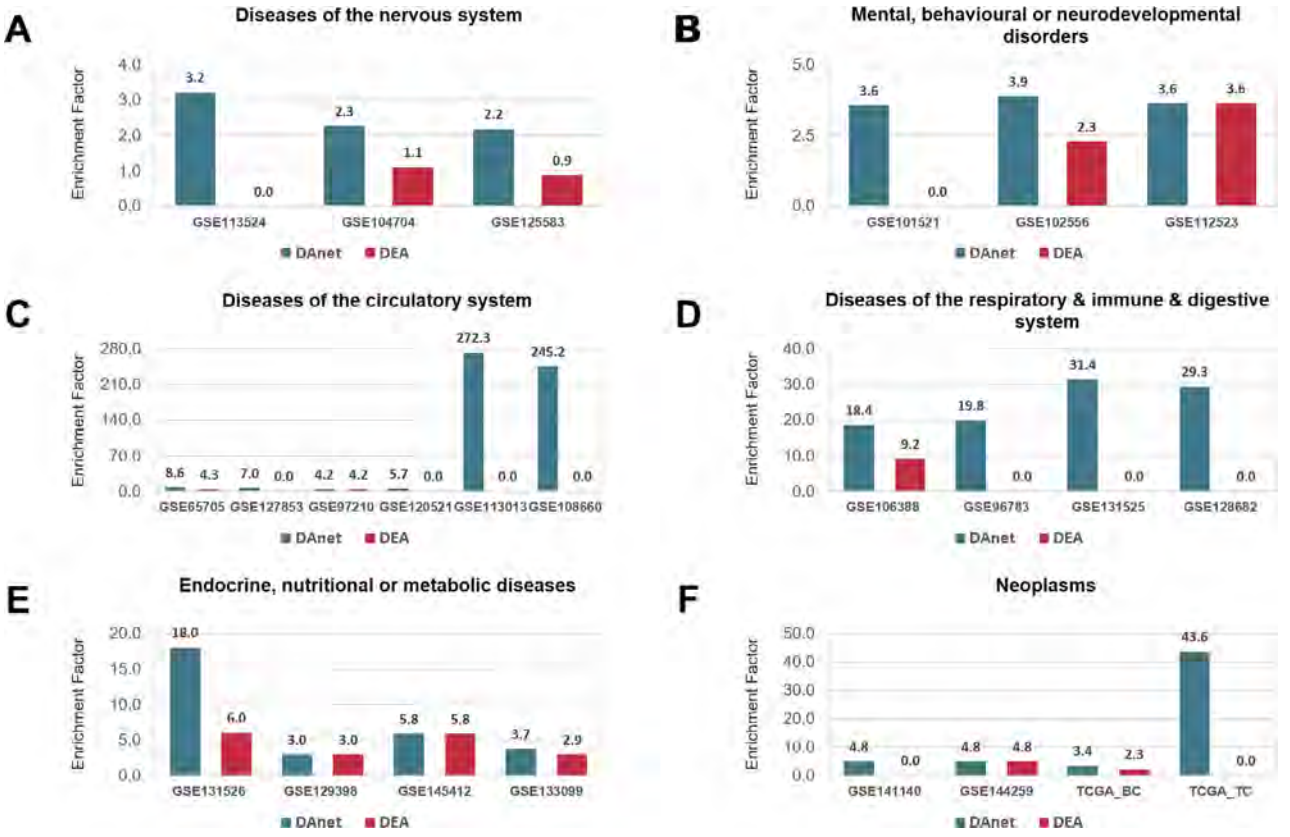


**Fig. 2.** Performance comparison between DAnet and DEA across the 24 benchmark datasets (shown in Table 1) based on the enrichment factor (EF), the EF represented the comparison between the concentration of the experimentally verified lncRNAs in the identification results of DAnet/DEA and the concentration in the entire lncRNAs expression.

for GSE106388). The Rate of DAnet was generally no less than DEA across 24 benchmark datasets. Moreover, among the 24 benchmark datasets, two datasets GSE97210 and GSE120521 from the atherosclerosis were collected from the microarray and RNA-Seq, respectively. We further compared the differences between the microarray and RNA-Seq data in terms of the originally detected lncRNAs, the potential disease-associated lncRNAs and the experimentally validated lncRNAs. As shown in the Supplementary Fig. S2, the total number of the originally detected lncRNAs for GSE97210 and GSE120521 was 10,347 and 10343, respectively. The number of lncRNAs detected by both GSE97210 and GSE120521 was 6836 (highlighted in blue and red lines). The number of potential disease-associated lncRNAs for GSE97210 and GSE120521 was 163 and 120, respectively. The number of shared lncRNAs was 111 (highlighted in green and red lines). In both GSE97210 and GSE120521, the experimentally validated lncRNA (CDKN2B-AS1) was identified via the DAnet. These findings indicate that both GSE97210 and GSE120521 are consistent in identifying the experimentally validated lncRNA.

Similarly, the EF was employed to assess the ability of DAnet and DEA about controlling the false characterization. As shown in Fig. 2, the EF of DAnet was differed greatly (from 2.2 for GSE125583 to 272.3 for GSE113013) and the EF of DEA was also varied (from 0.0 for 11 datasets to 9.2 for GSE106388). The EF of DAnet was generally no less than DEA of each dataset and all EFs of DAnet were greater than one.

### 3.2. Optimizing the $K_{CV}$ and CD parameters across the benchmark datasets

In order to identify more likely disease-associated lncRNAs, optimization procedure was performed to determine the optimal $K_{CV}$ and CD across the benchmark datasets. As shown in Fig. 3, the optimal $K_{CV}$ represented in red square was varied across the datasets (from 8 for TCGA-TC to 1000 for GSE102556), and the CV of experimentally verified disease-associated lncRNAs was generally higher. Table 2 shows the optimal $K_{CV}$ value across the datasets. Moreover, as shown in Supplementary Fig. S3, the optimal CD represented in red square was different across the datasets (from 5 kb for 13 datasets to 400 kb for GSE113524). Table 2 shows the optimal CD across the datasets. For six datasets (GSE127853, GSE97210, GSE113013, GSE108660, GSE141140, TCGA_TC), the CD was not available.

### 3.3. The function of lncRNA in disease characterized by DAnet

#### 3.3.1. KEGG enrichment analysis to character lncRNA function

Moreover, the co-expression network of lncRNAs and neighboring mRNAs was constructed under the optimal $K_{CV}$ and CD by WGCNA for each dataset. The network of module (contains the most genes with significant correlation) were displayed by Cytoscape [48]. Four networks are shown in Fig. 4 A-D as examples, the light-yellow square represented the lncRNA and the blue dot represented the co-expressed mRNA in the *cis*-lncRNA regulatory networks, red edge represented the association between disease-associated lncRNA and neighboring mRNA. Other 14 networks are shown in Supplementary Fig. S4. For each dataset, the KEGG enrichment analysis was performed to character lncRNA function via the co-expressed mRNAs. A chord diagram was dawn for illustrating the significantly enriched pathways across different datasets (Fig. 4 E). As shown in Fig. 4 E, the enriched pathways reported to be associated with the disease studied were indicated in blue lines, and other pathways were shown in grey lines. The statistical results of disease-related pathways in each dataset are

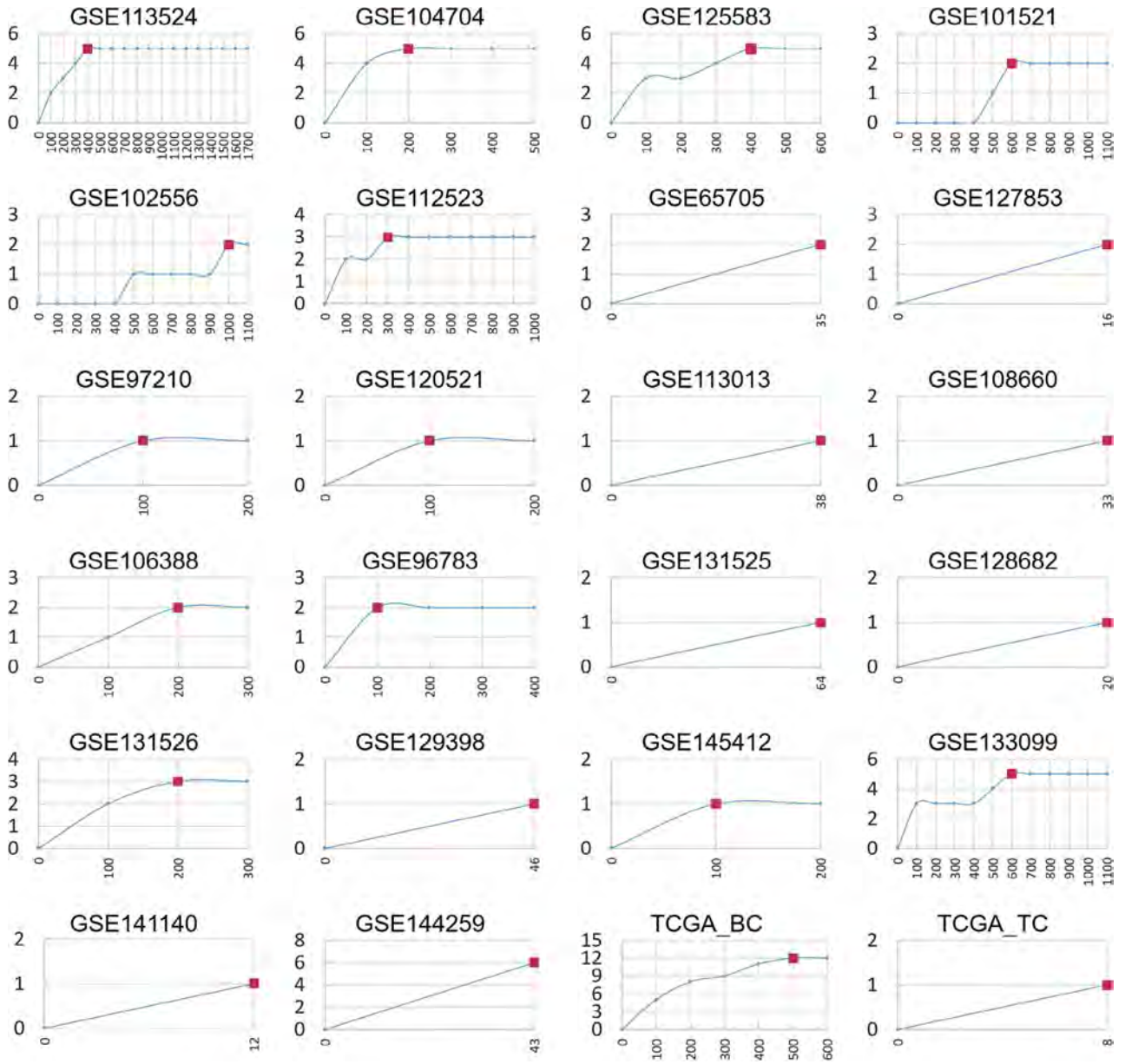


**Fig. 3.** Optimization for the $K_{CV}$ across these benchmark datasets. X axis: the top number of lncRNAs with the higher variabilities, Y axis: the number of experimental verified lncRNA ($N_{exp}$). When the number of lncRNA identified by SNPs ($N_{snp}$) was less than 100, the $K$ was equal to the $N_{snp}$, if else, the $K$ was from 100 to $N_{snp}$ with gradient of 100.

shown in Fig. 4 F. As shown, the percentage of disease-associated pathways were differed from 40% to 100% across datasets. The detail ed descriptions on relevance between disease and pathways are provided in Supplementary Table S2.

### *3.3.2. Association between lncRNAs identified by DAnet and the specific disease*

Finally, the relationships of lncRNAs and diseases were systemic manually searched. As illustrated in Fig. 5, 41 directly diseases-associated lncRNAs were identified for most diseases (blue lines). In particular, 13 lncRNAs were identified for Alzheimer disease (orange square, 8A20), three for major depressive disorder (brown square, 6A70), four for schizophrenia (brown square, 6A20), 12 for myocardial infarction (blue square, BA41), two for atherosclerosis (blue square, BD40), six for asthma (pink square, CA23), one for lupus erythematosus (purple square, 4A40), one for ulcerative colitis (turquoise square, DD71), five for obesity (yellow square, 5B81), six for type-2 diabetes mellitus (yellow square, 5A11), three for colorectal cancer (green square, 2B91), six for breast cancer (green square, 2C6Z). The detailed descriptions on relevance between lncRNAs and the specific disease are provided in Supplementary Table S3.

Meanwhile, as illustrated in Fig. 5, the lncRNAs (red dots) associated with multiple diseases were identified. Specifically, two lncRNAs (LINC-PINT, GAS5) were associated both with Alzheimer disease and type-2 diabetes mellitus [52–56], SOX2-OT was associated with Alzheimer disease and asthma [57,58], CCDC39 was associated with asthma and schizophrenia [59,60], HCP5 was associated with asthma and breast cancer [61,62], IFNG-AS1 was associated with asthma and ulcerative colitis [63,64], CDKN2B-AS1 was associated with five diseases including Alzheimer disease, myocardial infarction, atherosclerosis, type-2 diabetes mellitus, and breast cancer [65–70].

## 4. Discussion

Functional annotation of lncRNAs in diseases has attracted great attention for understanding disease etiology. In this study, we pro-

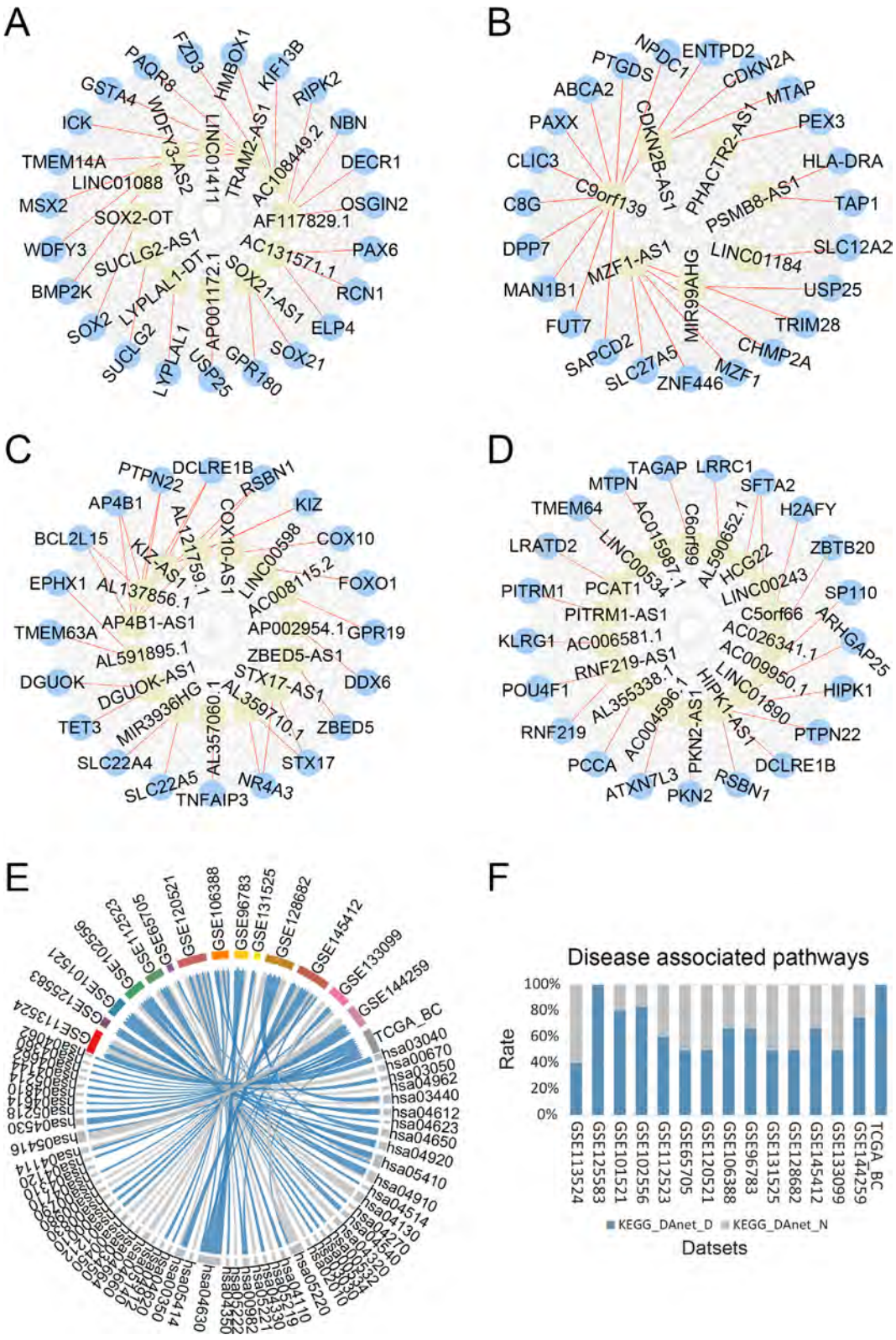


**Fig. 4.** The function of lncRNA in disease characterized by DAnet. A-D: co-expression network of module (contains the most genes with significant correlation) constructed by WGCNA for each dataset. A: GSE113524, B: GSE65705, C: GSE131525, D: GSE131526, green square: lncRNA, blue dot: mRNA. E: chord diagram of enriched pathways of 15 benchmark datasets ($p$ less than 0.05). F: the statistic of diseases-associated pathways.

posed a novel strategy termed DAnet by combining disease associations with *cis*-regulated network between lncRNAs and neighboring protein-coding genes for improving the functional annotation of lncRNAs. The strategy mainly consists of three procedures including: (1) identifying potential disease-associated lncRNAs based on disease-associated SNPs, (2) detecting more likely disease-associated lncRNAs based on expression variability, (3) developing *cis*-regulated networks between disease-associated lncRNAs and their neighboring protein-coding genes. To widen the scope of DAnet to other RNA-seq or Microarray data, the code

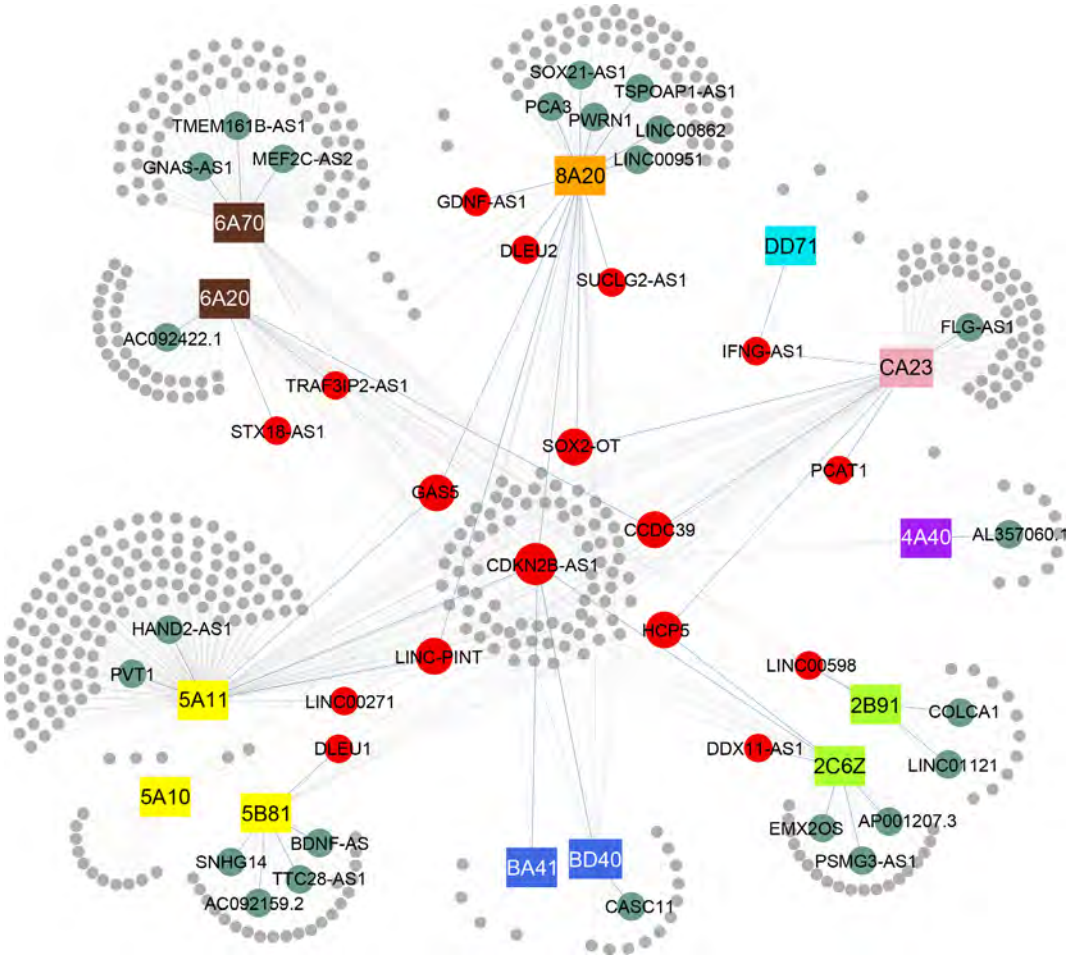


**Fig. 5.** Associations between lncRNAs identified by DAnet and the specific disease. The blue lines mean the reported associations between lncRNAs and diseases. The squares represent the type of diseases. The dots indicate lncRNAs identified by DAnet. Orange square: diseases of the nervous system; brown square: mental, behavioural or neurodevelopmental disorders; blue square: circulatory system disease; pink square: diseases of the respiratory system; purple square: diseases of the immune system; turquoise square: diseases of the digestive system; yellow square: endocrine, nutritional or metabolic diseases; green square: neoplasms; grey dot: lncRNA not reported in the studied disease; green dot: lncRNA associated with a single disease; red dot, lncRNA associated with multiple diseases.

of DAnet was provided in Supplementary Method S2. DAnet can be expected to identify the specific lncRNA function in the given disease.

Primarily, based on the analysis of 24 datasets involving 16 diseases, the Rate value of DAnet was overall higher than the DEA, which indicates that the performance of DAnet could be better than traditional differential expression-based analysis on identification of experimentally validated lncRNA. In addition, the EF of DAnet was overall higher than the DEA. All EFs of DAnet were higher than 1. These findings indicate the superior capacity of DAnet in controlling the false characterization of lncRNA function. Furthermore, during the optimization procedure for determining the optimal $K_{CV}$, we found that the experimentally verified disease-associated lncRNAs were generally with higher CV values. This finding is consistent with those reported by other investigators [16–18]. Under the optimal $K_{CV}$, the optimal CD was not available for these six datasets (GSE127853, GSE97210, GSE113013, GSE108660, GSE141140, TCGA_TC). This may be attributed to the effect of the small number of samples and the few numbers of lncRNAs/mRNAs in the co-expression analysis [71]. Finally, the KEGG enrichment results indicate most biological pathways identified by DAnet were associated with the corresponding disease (from 40% to 100%). And by DAnet, directly diseases-associated lncRNAs were identified for most diseases. Moreover, lncRNAs associated with multiple diseases were also identified.

## 5. Conclusions

A new strategy integrating disease associations was developed for obtaining the lower false discovery rate in functional annotation of lncRNAs. The analysis of 24 datasets involving 16 diseases, indicated that the performance of DAnet could be better than traditional differential expression-based on identification of experimentally validated lncRNA, and the most biological pathways identified by DAnet were associated with the studied diseases. This provides a way to study the function of lncRNA in diseases from another aspect. In sum, DAnet is expected to identify the specific lncRNA function in the given disease.

## Contributors

J.T. and Y.W. conceived the idea and supervised the work. Y.W., J.Z., and X.W., performed the research. Y.W., J.Z., X.W., Adu-Gyamfi E., L.Y., T.L., M.W., Y.D., and F.Z. prepared and analyzed the data. J.T. and Y.W. wrote manuscript. All authors reviewed and approved the final version of the manuscript.

## CRediT authorship contribution statement

**Yongheng Wang:** Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Jincheng Zhai:** Formal

analysis, Investigation. **Xianglu Wu:** Investigation, Visualization. **Enoch Appiah Adu-Gyamfi:** Validation, Writing – review & editing. **Lingping Yang:** Investigation, Validation. **Taihang Liu:** Validation. **Meijiao Wang:** Validation. **Yubin Ding:** Project administration. **Feng Zhu:** Conceptualization, Project administration. **Yingxiong Wang:** Conceptualization, Supervision, Funding acquisition. **Jing Tang:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision.

## 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.

## Acknowledgments

This work was supported by the National Key Research and Development Program of China (2018YFC1004400); the Natural Science Foundation of Chongqing (cstc2018jcyjAX0309) and the Outstanding Graduate Student Cultivation Program of Chongqing Medical University (BJRC201917).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.12.016.

## References

[1] Marchese FP, Raimondi I, Huarte M. The multidimensional mechanisms of long noncoding RNA function. Genome Biol 2017;18(1):206. https://doi.org/10.1186/s13059-017-1348-2.
[2] Kopp F, Mendell JT. Functional Classification and Experimental Dissection of Long Noncoding RNAs. Cell 2018;172(3):393–407. https://doi.org/10.1016/j.cell.2018.01.011.
[3] Chen G, Wang Z, Wang D, Qiu C, Liu M, et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res 2013;41 (Database issue):D983-986. Doi: 10.1093/nar/gks1099.
[4] Niknafs YS, Han S, Ma T, Speers C, Zhang C, Wilder-Romans K, et al. The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. Nat Commun 2016;7(1). https://doi.org/10.1038/ncomms12791.
[5] Micheletti R, Plaisance I, Abraham BJ, Sarre A, Ting C-C, Alexanian M, et al. The long noncoding RNA Wisper controls cardiac fibrosis and remodeling. Sci Transl Med 2017;9(395). https://doi.org/10.1126/scitranslmed.aai9118.
[6] Millan MJ. Linking deregulation of non-coding RNA to the core pathophysiology of Alzheimer's disease: An integrative review. Prog Neurobiol 2017;156:1–68. https://doi.org/10.1016/j.pneurobio.2017.03.004.
[7] Moran I, Akerman I, van de Bunt M, Xie R, Benazra M, et al. Human beta cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. Cell Metab 2012;16 (4):435–48. https://doi.org/10.1016/j.cmet.2012.08.010.
[8] Chen YG, Satpathy AT, Chang HY. Gene regulation in the immune system by long noncoding RNAs. Nat Immunol 2017;18(9):962–72. https://doi.org/10.1038/ni.3771.
[9] Alam T, Uludag M, Essack M, Salhi A, Ashoor H, et al. FARNA: knowledgebase of inferred functions of non-coding RNA transcripts. Nucleic Acids Res 2017;45 (5):2838-2848. https://doi.org/10.1093/nar/gkw973
[10] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf 2008;9(1):559. https://doi.org/10.1186/1471-2105-9-559.
[11] Jiao J, Shi B, Wang T, Fang Y, Cao T, et al. Characterization of long non-coding RNA and messenger RNA profiles in follicular fluid from mature and immature ovarian follicles of healthy women and women with polycystic ovary syndrome. Hum Reprod. 2018;33(9):1735-1748. https://doi.org/10.1093/humrep/dey255
[12] Khyzha N, Khor M, DiStefano PV, Wang L, Matic L, Hedin U, et al. Regulation of CCL2 expression in human vascular endothelial cells by a neighboring divergently transcribed long noncoding RNA. Proc Natl Acad Sci U S A. 2019;116(33):16410–9. https://doi.org/10.1073/pnas.1904108116.
[13] Yu J, Xuan Z, Feng X, Zou Q, Wang L. A novel collaborative filtering model for LncRNA-disease association prediction based on the Naive Bayesian classifier. BMC Bioinf 2019;20(1):396. https://doi.org/10.1186/s12859-019-2985-0.
[14] Bao Z, Yang Z, Huang Z, Zhou Y, Cui Q, et al. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res 2019;47(D1):D1034-D1037. https://doi.org/10.1093/nar/gky905
[15] Volders PJ, Anckaert J, Verheggen K, Nuytens J, Martens L, et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res 2019;47(D1):D135-D139. https://doi.org/10.1093/nar/gky1031
[16] Signal B, Gloss BS, Dinger ME. Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs. Trends Genet 2016;32(10):620–37. https://doi.org/10.1016/j.tig.2016.08.004.
[17] Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, et al. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. Genome Biol 2016;17(1). https://doi.org/10.1186/s13059-016-0873-8.
[18] Paraskevopoulou MD, Vlachos IS, Karagkouni D, Georgakilas G, Kanellos I, Vergoulis T, et al. DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. Nucleic Acids Res 2016;44(D1):D231–8. https://doi.org/10.1093/nar/gkv1270.
[19] Hua JT, Ahmed M, Guo H, Zhang Y, Chen S, Soares F, et al. Risk SNP-Mediated Promoter-Enhancer Switching Drives Prostate Cancer through lncRNA PCAT19. Cell 2018;174(3):564–575.e18. https://doi.org/10.1016/j.cell.2018.06.014.
[20] Castellanos-Rubio A, Ghosh S. Disease-Associated SNPs in Inflammation-Related lncRNAs. Front Immunol 2019;10:420. https://doi.org/10.3389/fimmu.2019.00420.
[21] Guo H, Ahmed M, Zhang F, Yao CQ, Li SiDe, Liang Yi, et al. Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to prostate cancer. Nat Genet 2016;48(10):1142–50. https://doi.org/10.1038/ng.3637.
[22] Ecker S, Pancaldi V, Rico D, Valencia A. Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia. Genome Med 2015;7(1):8. https://doi.org/10.1186/s13073-014-0125-z.
[23] Cho SW, Xu J, Sun R, Mumbach MR, Carter AC, Chen YG, et al. Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. Cell 2018;173(6):1398–1412.e22. https://doi.org/10.1016/j.cell.2018.03.068.
[24] Suzuki S, Shaw G, Renfree MB. Identification of a novel antisense noncoding RNA, ALID, transcribed from the putative imprinting control region of marsupial IGF2R. Epigenet Chromatin 2018;11(1):55. https://doi.org/10.1186/s13072-018-0227-8.
[25] Jiang W, Liu Y, Liu R, Zhang K, Zhang Y. The lncRNA DEANR1 facilitates human endoderm differentiation by activating FOXA2 expression. Cell Rep 2015;11 (1):137–48. https://doi.org/10.1016/j.celrep.2015.03.008.
[26] Brazao TF, Johnson JS, Muller J, Heger A, Ponting CP, et al. Long noncoding RNAs in B-cell development and activation. Blood 2016;128(7):e10–19. https://doi.org/10.1182/blood-2015-11-680843.
[27] 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.
[28] Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell 2018;173(2):283–5. https://doi.org/10.1016/j.cell.2018.03.042.
[29] Ma L, Li A, Zou D, Xu X, Xia L, et al. LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs. Nucleic Acids Res. 2015;43(Database issue):D187-192. https://doi.org/10.1093/nar/gku1167
[30] Cheng L, Wang P, Tian R, Wang S, Guo Q, et al. LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res 2019;47(D1):D140-D144. https://doi.org/10.1093/nar/gky1051
[31] Gao Y, Wang P, Wang Y, Ma X, Zhi H, et al. Lnc2Cancer v2.0: updated database of experimentally supported long non-coding RNAs in human cancers. Nucleic Acids Res 2019;47(D1):D1028-D1033. https://doi.org/10.1093/nar/gky1096
[32] Zhou B, Zhao H, Yu J, Guo C, Dou X, et al. EVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput experiments. Nucleic Acids Res 2018;46(D1):D100-D105. https://doi.org/10.1093/nar/gkx677
[33] Eicher JD, Landowski C, Stackhouse B, Sloan A, Chen W, et al. GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. Nucleic Acids Res. 2015;43(Database issue):D799-804. https://doi.org/10.1093/nar/gku1202
[34] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019;47(D1):D1005-D1012. https://doi.org/10.1093/nar/gky1120
[35] Li MJ, Liu Z, Wang P, Wong MP, Nelson MR, Kocher J-P, et al. GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res 2016;44(D1):D869–76. https://doi.org/10.1093/nar/gkv1317.
[36] Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766-D773. https://doi.org/10.1093/nar/gky955
[37] Liao Q, Xiao H, Bu D, Xie C, Miao R, Luo H, et al. ncFANs: a web server for functional annotation of long non-coding RNAs. Nucleic Acids Res 2011;39 (suppl):W118–24.
[38] Ros G, Pegoraro S, De Angelis P, Sgarra R, Zucchelli S, Gustincich S, et al. HMGA2 Antisense Long Non-coding RNAs as New Players in the Regulation of HMGA2 Expression and Pancreatic Cancer Promotion. Front Oncol 2019;9. https://doi.org/10.3389/fonc.2019.01526.
[39] Wang M, Yuan D, Tu L, Gao W, He Y, et al. Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.). New Phytol 2015;207(4):1181–97. https://doi.org/10.1111/nph.13429.
[40] Cabili MN, Dunagin MC, McClanahan PD, Biaesch A, Padovan-Merhar O, Regev A, et al. Localization and abundance analysis of human lncRNAs at single-cell

and single-molecule resolution. Genome Biol 2015;16(1). https://doi.org/10.1186/s13059-015-0586-4.
[41] Werner MS, Sullivan MA, Shah RN, Nadadur RD, Grzybowski AT, Galat V, et al. Chromatin-enriched lncRNAs can act as cell-type specific activators of proximal gene transcription. Nat Struct Mol Biol 2017;24(7):596–603. https://doi.org/10.1038/nsmb.3424.
[42] Teimuri S, Hosseini A, Rezaenasab A, Ghaedi K, Ghoveud E, Etemadifar M, et al. Integrative Analysis of lncRNAs in Th17 Cell Lineage to Discover New Potential Biomarkers and Therapeutic Targets in Autoimmune Diseases. Mol Ther Nucleic Acids 2018;12:393–404. https://doi.org/10.1016/j.omtn.2018.05.022.
[43] Li S, Yu X, Lei N, Cheng Z, Zhao P, He Y, et al. Genome-wide identification and functional prediction of cold and/or drought-responsive lncRNAs in cassava. Sci Rep 2017;7(1). https://doi.org/10.1038/srep45981.
[44] Wang X, Yang C, Guo F, Zhang Y, Ju Z, Jiang Q, et al. Integrated analysis of mRNAs and long noncoding RNAs in the semen from Holstein bulls with high and low sperm motility. Sci Rep 2019;9(1). https://doi.org/10.1038/s41598-018-38462-x.
[45] Schultz BM, Gallicio GA, Cesaroni M, Lupey LN, Engel N. Enhancers compete with a long non-coding RNA for regulation of the Kcnq1 domain. Nucleic Acids Res 2015;43(2):745–59. https://doi.org/10.1093/nar/gku1324.
[46] Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, et al. Long noncoding RNAs with enhancer-like function in human cells. Cell 2010;143(1):46–58.
[47] Pyfrom SC, Luo H, Payton JE. PLAIDOH: a novel method for functional prediction of long non-coding RNAs identifies cancer-specific LncRNA activities. BMC Genomics 2019;20(1):137. https://doi.org/10.1186/s12864-019-5497-4.
[48] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498–504. https://doi.org/10.1101/gr.1239303.
[49] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45(D1):D353–61. https://doi.org/10.1093/nar/gkw1092.
[50] Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular visualization in R. Bioinformatics 2014;30(19):2811–2. https://doi.org/10.1093/bioinformatics/btu393.
[51] Hong J, Luo Y, Zhang Y, Ying J, Xue W, et al. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. Brief Bioinform 2020;21(4):1437-1447. https://doi.org/10.1093/bib/bbz081
[52] Simchovitz A, Hanan M, Yayon N, Lee S, Bennett ER, Greenberg DS, et al. A lncRNA survey finds increases in neuroprotective LINC-PINT in Parkinson's disease substantia nigra. Aging Cell 2020;19(3). https://doi.org/10.1111/acel.v19.310.1111/acel.13115.
[53] Zha T, Su F, Liu X, Yang C, Liu L. Role of Long Non-Coding RNA (LncRNA) LINC-PINT Downregulation in Cardiomyopathy and Retinopathy Progression Among Patients with Type 2 Diabetes. Med Sci Monit 2019;25:8509–14. , https://doi.org/10.12659/MSM.918358.
[54] Xu W, Zhang L, Geng Yu, Liu Ye, Zhang N. Long noncoding RNA GAS5 promotes microglial inflammatory response in Parkinson's disease by regulating NLRP3 pathway through sponging miR-223-3p. Int Immunopharmacol 2020;85:106614. https://doi.org/10.1016/j.intimp.2020.106614.
[55] Carter G, Miladinovic B, Patel AA, Deland L, Mastorides S, Patel NA. Circulating long noncoding RNA GAS5 levels are correlated to prevalence of type 2 diabetes mellitus. BBA Clin 2015;4:102–7. https://doi.org/10.1016/j.bbacli.2015.09.001.
[56] Li Z, Yu Z, Meng X, Zhou S, Xiao S, et al. Long noncoding RNA GAS5 impairs the proliferation and invasion of endometrial carcinoma induced by high glucose via targeting miR-222-3p/p27. Am J Transl Res 2019;11(4):2413–21.
[57] Arisi I, D'Onofrio M, Brandi R, Felsani A, Capsoni S, Drovandi G, et al. Gene expression biomarkers in the brain of a mouse model for Alzheimer's disease: mining of microarray data by logic classification and feature selection. J Alzheimers Dis 2011;24(4):721–38. https://doi.org/10.3233/JAD-2011-101881.
[58] Saghaeian Jazi M, Samaei NM, Mowla SJ, Arefnezhad B, Kouhsar M. SOX2OT knockdown derived changes in mitotic regulatory gene network of cancer cells. Cancer Cell Int 2018;18:129. https://doi.org/10.1186/s12935-018-0618-8.
[59] Guo Z, Chen W, Wang L, Qian L. Clinical and Genetic Spectrum of Children with Primary Ciliary Dyskinesia in China. J Pediatr 2020;225:157–165.e5. https://doi.org/10.1016/j.jpeds.2020.05.052.
[60] . Nature 2014;511(7510):421–7. https://doi.org/10.1038/nature13595.
[61] Choi S, Park HS, Cheon MS, Lee K. Expression profile analysis of human peripheral blood mononuclear cells in response to aspirin. Arch Immunol Ther Exp (Warsz) 2005;53(2):151–8.
[62] Wang L, Luan T, Zhou S, Lin J, Yang Y, Liu W, et al. LncRNA HCP5 promotes triple negative breast cancer progression as a ceRNA to regulate BIRC3 by sponging miR-219a-5p. Cancer Med. 2019;8(9):4389–403. https://doi.org/10.1002/cam4.2335.
[63] Bérubé J-C, Gaudreault N, Lavoie-Charland E, Sbarra L, Henry C, Madore A-M, et al. Identification of Susceptibility Genes of Adult Asthma in French Canadian Women. Can Respir J. 2016;2016:1–12. https://doi.org/10.1155/2016/3564341.
[64] Padua D, Mahurkar-Joshi S, Law IKM, Polytarchou C, Vu JP, Pisegna JR, et al. A long noncoding RNA signature for ulcerative colitis identifies IFNG-AS1 as an enhancer of inflammation. Am J Physiol Gastrointest Liver Physiol 2016;311(3):G446–57. https://doi.org/10.1152/ajpgi.00212.2016.
[65] Züchner S, Gilbert JR, Martin ER, Leon-Guerrero CR, Xu P-T, Browning C, et al. Linkage and association study of late-onset Alzheimer disease families linked to 9p21.3. Ann Hum Genet 2008;72(6):725–31. https://doi.org/10.1111/j.1469-1809.2008.00474.x.
[66] AbdulAzeez S, Al-Nafie A, Al-Shehri A, Borgio J, Baranova E, Al-Madan M, et al. Intronic Polymorphisms in the CDKN2B-AS1 Gene Are Strongly Associated with the Risk of Myocardial Infarction and Coronary Artery Disease in the Saudi Population. Int J Mol Sci 2016;17(3):395. https://doi.org/10.3390/ijms17030395.
[67] Li Y, Zhang D, Zhang Y, Xu X, Bi L, Zhang M, et al. Association of lncRNA polymorphisms with triglyceride and total cholesterol levels among myocardial infarction patients in Chinese population. Gene 2020;724:143684. https://doi.org/10.1016/j.gene.2019.02.085.
[68] Li H, Han S, Sun Q, Yao Ye, Li S, Yuan C, et al. Long non-coding RNA CDKN2B-AS1 reduces inflammatory response and promotes cholesterol efflux in atherosclerosis by inhibiting ADAM10 expression. Aging (Albany NY). 2019;11(6):1695–715.
[69] Hubberten M, Bochenek G, Chen H, Häsler R, Wiehe R, Rosenstiel P, et al. Linear isoforms of the long noncoding RNA CDKN2B-AS1 regulate the c-myc-enhancer binding factor RBMS1. Eur J Hum Genet 2019;27(1):80–9. https://doi.org/10.1038/s41431-018-0210-7.
[70] Bozgeyik E, Saadat KASM, Arman K, Bozgeyik I, Ikeda M-A. Enhanced E2F1 activity increases invasive and proliferative activity of breast cancer cells through non-coding RNA CDKN2B-AS1. Meta. Gene 2020;24:100691. https://doi.org/10.1016/j.mgene.2020.100691.
[71] van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform 2018;19(4):575–92. https://doi.org/10.1093/bib/bbw139.
[72] Ivashko-Pachima Y, Hadar A, Grigg I, Korenková V, Kapitansky O, Karmon G, et al. Discovery of autism/intellectual disability somatic mutations in Alzheimer's brains: mutated ADNP cytoskeletal impairments and repair as a case study. Mol Psychiatry 2021;26(5):1619–33. https://doi.org/10.1038/s41380-019-0563-5.
[73] Nativio R, Donahue G, Berson A, Lan Y, Amlie-Wolf A, Tuzer F, et al. Dysregulation of the epigenetic landscape of normal aging in Alzheimer's disease. Nat Neurosci 2018;21(4):497–505. https://doi.org/10.1038/s41593-018-0101-9.
[74] Srinivasan K, Friedman BA, Etxeberria A, Huntley MA, van der Brug MP, Foreman O, et al. Alzheimer's Patient Microglia Exhibit Enhanced Aging and Unique Transcriptional Activation. Cell Rep. 2020;31(13):107843. https://doi.org/10.1016/j.celrep.2020.107843.
[75] Pantazatos SP, Huang Y-Y, Rosoklija GB, Dwork AJ, Arango V, Mann JJ. Whole-transcriptome brain expression and exon-usage profiling in major depression and suicide: evidence for altered glial, endothelial and ATPase activity. Mol Psychiatry 2017;22(5):760–73. https://doi.org/10.1038/mp.2016.130.
[76] Labonté B, Engmann O, Purushothaman I, Menard C, Wang J, Tan C, et al. Sex-specific transcriptional signatures in human depression. Nat Med 2017;23(9):1102–11. https://doi.org/10.1038/nm.4386.
[77] Pai S, Li P, Killinger B, Marshall L, Jia P, Liao Ji, et al. Differential methylation of enhancer at IGF2 is associated with abnormal dopamine synthesis in major psychosis. Nat Commun 2019;10(1). https://doi.org/10.1038/s41467-019-09786-7.
[78] Eicher JD, Wakabayashi Y, Vitseva O, Esa N, Yang Y, Zhu J, et al. Characterization of the platelet transcriptome by RNA sequencing in patients with acute myocardial infarction. Platelets 2016;27(3):230–9. https://doi.org/10.3109/09537104.2015.1083543.
[79] Li J, Wu J, Zhang M, Zheng Y. Dynamic changes of innate lymphoid cells in acute ST-segment elevation myocardial infarction and its association with clinical outcomes. Sci Rep 2020;10(1):5099. https://doi.org/10.1038/s41598-020-61903-5.
[80] Hu YW, Guo FX, Xu YJ, Li P, Lu ZF, et al. Long noncoding RNA NEXN-AS1 mitigates atherosclerosis by regulating the actin-binding protein NEXN. J Clin Invest 2019; 129(3):1115-1128. https://doi.org/10.1172/JCI98230
[81] Mahmoud AD, Ballantyne MD, Miscianinov V, Pinel K, Hung J, Scanlon JP, et al. The Human-Specific and Smooth Muscle Cell-Enriched LncRNA SMILR Promotes Proliferation by Regulating Mitotic CENPF mRNA and Drives Cell-Cycle Progression Which Can Be Targeted to Limit Vascular Remodeling. Circ Res 2019;125(5):535–51. https://doi.org/10.1161/CIRCRESAHA.119.314876.
[82] Ravi A, Koster J, Dijkhuis A, Bal SM, Sabogal Piñeros YS, Bonta PI, et al. Interferon-induced epithelial response to rhinovirus 16 in asthma relates to inflammation and FEV1. J Allergy Clin Immunol 2019;143(1):442–447.e10. https://doi.org/10.1016/j.jaci.2018.09.016.
[83] Altman MC, Whalen E, Togias A, O'Connor GT, Bacharier LB, Bloomberg GR, et al. Allergen-induced activation of natural killer cells represents an early-life immune response in the development of allergic asthma. J Allergy Clin Immunol 2018;142(6):1856–66. https://doi.org/10.1016/j.jaci.2018.02.019.
[84] Fenton CG, Taman H, Florholmen J, Sorbye SW, Paulssen RH. Transcriptional Signatures That Define Ulcerative Colitis in Remission. Inflamm Bowel Dis 2020. https://doi.org/10.1093/ibd/izaa075
[85] Speake C, Skinner SO, Berel D, Whalen E, Dufort MJ, Young WC, et al. A composite immune signature parallels disease progression across T1D subjects. JCI Insight 2019;4(23). https://doi.org/10.1172/jci.insight.12691710.1172/jci.insight.126917DS1.

[86] Herring BP, Chen M, Mihaylov P, Hoggatt AM, Gupta A, Nakeeb A, et al. Transcriptome profiling reveals significant changes in the gastric muscularis externa with obesity that partially overlap those that occur with idiopathic gastroparesis. BMC Med Genomics 2019;12(1). https://doi.org/10.1186/s12920-019-0550-3.

[87] Paczkowska-Abdulsalam M, Niemira M, Bielska A, Szałkowska A, Raczkowska BA, Junttila S, et al. Evaluation of Transcriptomic Regulations behind Metabolic Syndrome in Obese and Lean Subjects. Int J Mol Sci 2020;21(4):1455. https://doi.org/10.3390/ijms21041455.

[88] Shu Yi, Wang Yi, Lv W-Q, Peng D-Y, Li J, Zhang H, et al. ARRB1-Promoted NOTCH1 Degradation Is Suppressed by OncomiR miR-223 in T-cell Acute Lymphoblastic Leukemia. Cancer Res 2020;80(5):988–98. https://doi.org/10.1158/0008-5472.CAN-19-1471.

[89] Ji Q, Zhou L, Sui H, Yang L, Wu X, Song Q, et al. Primary tumors release ITGBL1-rich extracellular vesicles to promote distal metastatic tumor growth through fibroblast-niche formation. Nat Commun 2020;11(1). https://doi.org/10.1038/s41467-020-14869-x.