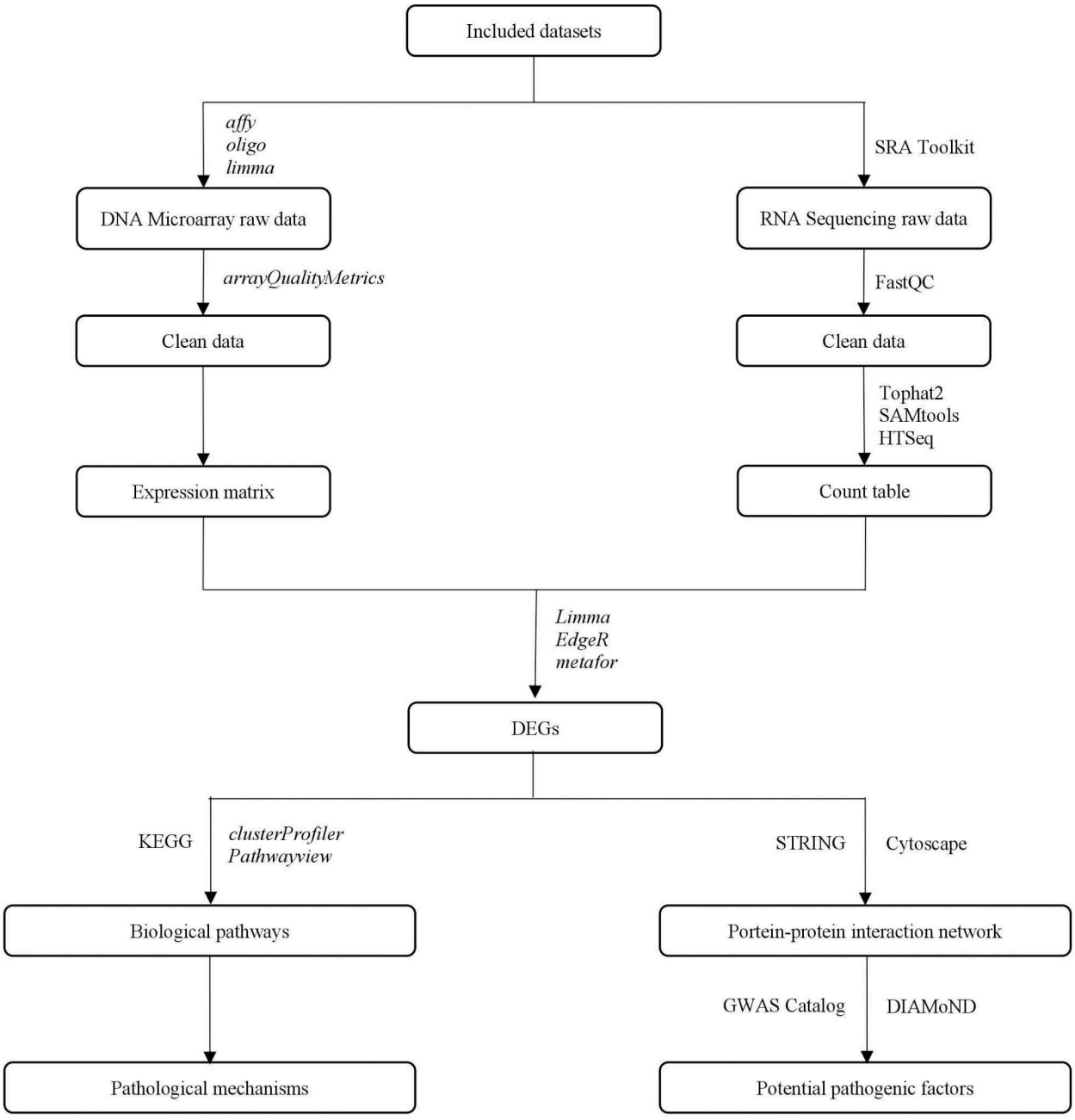
Introduction and whole summary.

This study aims to identify potential pathogens factors based on a meta-analysis of transcriptome data from different brain regions of AD Experiment:

AD is characterized by the distribution of amyloid plaques and neurofibrillary tangles, which are composed of Aß and hyperphosphorylated tau proteins.
→ Generations of Aß from amyloid precursor protein(APP).
→ Aß are of 2 types- Aß40 & Aß41
→ Aß monomer is highly binding affinity, so it combines with oligomers, fibrils, and plaques (Aß oligomer is recognized as major neurotoxic species.
→ Aß is directly neurotoxic and induces a series of cellular responses eg, it can alter Ca2+ homeostasis, synaptic function in neuron
→ Aß also activates protein kinase to phosphorylate tau proteins, resulting in the formation of neurofibrillary tangle.
→ The amyloid cascade hypothesis proposed in 1992 depicts Aß as a causative factor in AD development. supporting this hypothesis as genetic-based mutations related to Aß formation (APP and PSEN) and clearance (APOEY) induce early-onset AD. → The drug development for the treatment of AD has been guided by the amyloid cascade hypothesis for the past 2 decades.
→ The discovery of AD pathogens factors can be revealed by transcriptomic profiling approach (ie, DNA microarray RNA seq) to identify DEG in an Aß subject compared to a healthy subject.
→ Integrations of these studies based on meta-analysis, therefore, resolve inconclusive results and obtain a generalization.
→ Moradifard et al. (2018) conducted a meta-analysis using the R package Robust Rook Aggy to identify the statistically significant DEG in AD across six microarray datasets.

Title	Authors (publication year)	Database	Bioinformatic tools	Pathogenic factors
Integrated Identification of Key Genes and Pathways in Alzheimer's Disease via Comprehensive Bioinformatical Analyses (Yan et al., 2019)	Yan et al. (2019)	GEO, KEGG, Reactome, STRING, Wikipathway	Platform: Morpheus, RBPDB, UCSC; Software: Cytoscape, ClueGO, Gluepedia, Graphpad Prism, MCODE	*BDNF*, *CACNA1A*, *CALB*, *CD44*, *CDC42*, *OXT*, *PDYN*, *TAC1*, *TH*, *VEGFA*
Systematic Analysis and Biomarker Study for Alzheimer's Disease (Li X. et al., 2018)	Li X. et al. (2018)	GEO, International Genomics of Alzheimer's Project	R package: *affy*, *glmnet*, *limma*, *MASS*, *PRROC*, *ROCR*; Software: Ingenuity Pathway Analysis, MAGMA	*NDUFA1*, *MRPL51*, *RPL36AL*
Network-based approach to identify molecular signatures and therapeutic agents in Alzheimer's Disease (Rahman et al., 2019)	Rahman et al. (2019)	CMap, GEO, JASPAR, miRTarBase, STRING, TarBase	Platform: DAVID, GEO2R; Software: CytoHubba, Cytoscape, MCODE	*AR*, *CREBBP*, *E2F1*, *FOXC1*, *FOXL1*, *GATA2*, *JUN*, *NFIC*, *PPARG*, *RAC1*, *RPL12*, *RPL15*, *RPS11*, *RPS6*, *SMAD3*, *SRF*, *UBA52*, *UBC*, *USF2*, *YY1*
Alzheimer's Disease Master Regulators Analysis: Search for Potential Molecular Targets and Drug Repositioning Candidates (Vargas et al., 2018)	Vargas et al. (2018)	CMap, GEO	Algorithm: Algorithm for the Reconstruction of Accurate Cellular Networks, two-tail gene set enrichment analysis; R package: *ggplot2*, *RedeR*, *RTN*	*ATF2*, *CNOT7*, *CSRNP2*, *PARK2*, *SLC30A9*, *TSC22D1*
Condition-specific Gene Co-expression Network Mining Identifies Key Pathways and Regulators in the Brain Tissue of Alzheimer's Disease Patients (Xiang et al., 2018)	Xiang et al. (2018)	Allen Brain Institute, GEO	Algorithm: Local maximized Quasi-Clique Merger; Platform: REViGO; R package: *Affy*, *Enrichr*, *lmQCM*	*FOS, JUN, MEF2A, MIB2, PCBP1, SMARCA2, SP1, STAT1, TEAD4, ZFHX3, ZNF281*
Analytical Strategy to Prioritize Alzheimer's Disease Candidate Genes in Gene Regulatory Networks Using Public Expression Data (Kawalia et al., 2017)	Kawalia et al. (2017)	ArrayExpress, CPDB, ENSEMBL, GEO, GWAScatalog, GWASCentral, GWASdb, KEGG, NeuroTransDB, RegulomeDB	Algorithm: BC3Net10; R package: *affy*, *arrayQualityMetrics*, *bc3net*, *limma*; Platform: HaploReg, SCAIView	*AP2A2, ARAP3, ATP2A3, ATP2B4, HLA-C, HLA-F, ITPR2, RAB11FIP4, STX2*
The Bioinformatic Analysis of the Dysregulated Genes and MicroRNAs in Entorhinal Cortex, Hippocampus, and Blood for Alzheimer's Disease (Pang et al., 2017)	Pang et al. (2017)	CMap, GEO, KEGG	R package: *affy*, *edgeR*, *limma*, *WGCNA*; Platform: DAVID; Software: CytoHubba, Cytoscape, GSEA	*CTSD*, *VCAM1*
A Systematic Integrated Analysis of Brain Expression Profiles Reveals *YAP1* and Other Prioritized Hub Genes as Important Upstream Regulators in Alzheimer's Disease (Xu et al., 2018)	Xu et al. (2018)	GEO	R package: *in silico Merging*, *limma*, *WGCNA*	*YAP1*
Network Topology Analysis of Post-Mortem Brain Microarrays Identifies More Alzheimer's Related Genes and MicroRNAs and Points to Novel Routes for Fighting with the Disease (Chandrasekaran and Bonchev, 2016)	Chandrasekaran and Bonchev (2016)	ArrayExpress, GEO, OMIN, ResNet	Algorithm: Robust Multiarray Average approach, empirical Bayes method; Platform: DAVID; Software: Pajek, Pathway Studio	*CD4*, *DCN*, *IL8*
A Systematic Investigation into Aging Related Genes in Brain and Their Relationship with Alzheimer's Disease (Meng et al., 2016)	Meng et al. (2016)	Biocarta, DisGeNet, GEO, GenAge, GenMapP, MetaBase	Algorithm: Condition-specific target prediction; Platform: DAVID, oPOSSUM; R package: *WGCNA*; Software: Ingenuity Pathway Analysis	*ESR1*, *SOX2*, *SP1*
Identification of Differentially Expressed Genes through Integrated Study of Alzheimer's Disease Affected Brain Regions (Puthiyedth et al., 2016)	Puthiyedth et al. (2016)	GEO	Algorithm: (α,β)-k Feature Set approach, Minimum Description Length Principle; R package: *GeneMeta*, *RankProd*; Software: Expression Analysis Systematic Explorer	*FGF*, *GPHN*, *INFAR2*, *LARGE*, *PSMB2*, *PSMD14*, *PTMA*, *RAB2A*, *RPL15*, *RWDD2A*, *SEMA4C*, *WNK1*
Meta-Analysis of Transcriptome Data Related to Hippocampus Biopsies and iPSC-Derived Neuronal Cells from Alzheimer's Disease Patients Reveals an Association with FOXA1 and FOXA2 Gene Regulatory Networks (Wruck et al., 2016)	Wruck et al. (2016)	GEO, KEGG	Platform: oPOSSUM; R package: *affy*, *Gostats*, *lumi*, *oligo*	*FOXA1*, *FOXA2*
A Computational Framework for the Prioritization of Disease-gene Candidates (Browne et al., 2015)	Browne et al. (2015)	BIND, BioGPS, BioGRID, DIP, GEO, HPRD, InACT, MINT, PDB	Algorithm: Average of Pearson correlation coefficients; Platform: hORFeome; R package: *GOSemSim*; Software: Cytoscape, Significance	*CARD9*, *FHL3*, *KRT38*, *LZTS2*, *MID2*, *MTUS2*, *REL*, *TFCP2*, *TRAF1*


Included datasets
affy
oligo
limma
SRA Toolkit
DNA Microarray raw data
RNA Sequencing raw data
arrayQualityMetrics
FastQC
Clean data
Clean data
Tophat2
SAMtools
HTSeq
Expression matrix
Count table
Limma
EdgeR
metafor
DEGs
KEGG
clusterProfiler
Pathwayview
STRING
Cytoscape
Biological pathways
Portein-protein interaction network
GWAS Catalog
DIAMoND
Pathological mechanisms
Potential pathogenic factors

Database Search, Database Selection and Data Processing

- Datasets collected from GEO and ArrayExpress. Raw DNA microarray and RNA-Seq data reporting the transcriptomes of brain regions from AD and healthy subjects were included
- Each dataset was processed through background correction, quantile normalization, and log2 -transformation of the averaged expression value of duplicate probes, using the affy, oligo or limma package
- Quality of each array, for each dataset was assessed using arrayQualityMetrics package. Limma Package used to determine the gene expression ratios between AD and healthy subjects
- The DEGs were annotated using the gene annotation table provided by GEO and ArrayExpress. The probes with missing gene symbols were removed and The probe with the highest expression value was kept when multiple probes were annotated with the same gene symbol on the same array.
- The raw sequence data were converted to FASTQ format and Quality control was conducted using FastQC. The cleaner FASTQ files were aligned to the *Homo sapiens* reference genome, GRCh38.94
- Differential Expression was performed using the edgeR.

Meta-Analysis, Biological Enrichment Analysis, and Subgroup Analysis

- The list of DEGs and their (upregulated or downregulated) states from each study was processed using the meta package.
- Meta-analysis was conducted under a random-effects model, and their outcomes were logORs with *P*-values
- For each gene in the i-th study, the effect (θi) based on the numbers of (upregulated or downregulated) events in both AD and control samples was first calculated, then the overall effect was computed according to formula∑Wiθi/∑Wi, where wi is the weight and is equal to 1/vi, where vi is the sample variance.
- Genes with logORs above or below 0 were considered upregulated or downregulated resp. P-values were adjusted and less than 0.05 were regarded statistically significant
- After meta-analysis, the statistically significant DEGs were used to perform a biological enrichment analysis to identify biological pathways among them which was conducted by the hypergeometric test using the *cluster profiler.* The pathways with FDR-adjusted *P* -values less than 0.05 were considered statistically significant
- The DEGs were split into different subgroups based on which brain regions they were from, to compare expression profiles among brain regions.

Protein-Protein Interaction Network and Potential Pathogenic Factors

1)After meta-analysis, statistically significant DEGs were used to construct a PPIN(Protein-Protein Interaction Network) based on the data from STRING.Only interactions with the highest confidence (0.9) were kept.The PPIN was visualized using Cytoscape.
2)AD seed genes were retrieved from the GWAS Catalog (Welter et al., 2014)5 using the keyword “AD” and the following selection criteria: a significance p-Value cutoff of no more than 1∗10–8.
3)The PPIN information (e.g., edge list in the co-expression network) and AD seed genes were put into the DIAMOnD algorithm.
4)In each iteration, until the stopping condition was satisfied, the node with the lowest connectivity P-value was treated as the most significantly connected node for output. The nodes within the module that had a high degree (>95th percentile) were treated as potential pathogenic factors.

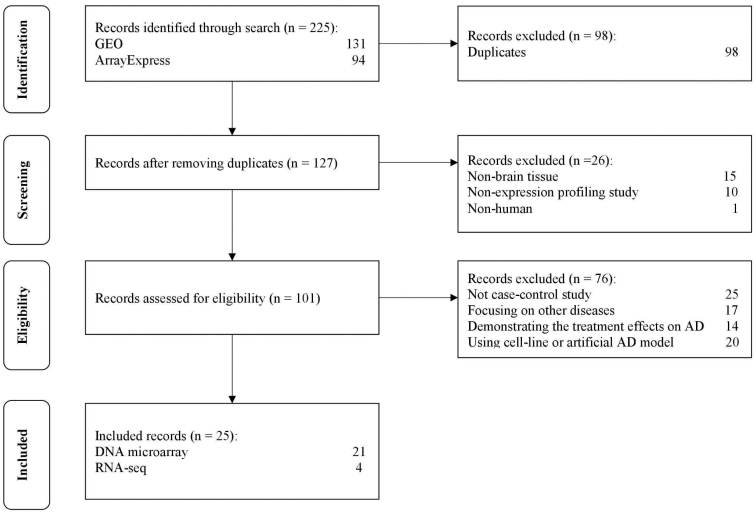
Result:-

Selecting data set:-

From GEO and ArrayExpress, 225 datasets were initially found. 25 datasets (21 for microarray and 4 for RNA-Seq) fulfilled the eligibility requirements after duplicates and datasets that met the exclusion criteria were removed. Following is the list of 225 datasets, **among the included datasets, the hippocampus was the most extensively studied brain region.**

Types	Dataset_ID	Brain tissue	Disease stage	Platform	Number of samples	Number of samples omitted
DNA microarray	E-MEXP-2280	Medial temporal lobe	Braak stage VI	Affymetrix Human Genome U133 Plus 2.0 Array	12	0
	GSE110226	Choroid plexus	Braak stage III to VI	Rosetta/Merck Human RSTA Custom Affymetrix 2.0 microarray	13	3
	GSE12685	Frontal cortex	MMSE 21 to 27	Affymetrix Human Genome U133A Array	14	3
	GSE1297	Hippocampus	Braak stage at III to VI	Affymetrix Human Genome U133A Array	31	3
	GSE16759	Parietal lobe cortex	Braak stage at V to VI	Affymetrix Human Genome U133 Plus 2.0 Array	8	1
	GSE26927	Entorhinal cortex	–	Illumina humanRef-8 v2.0 expression beadchip	18	4
	GSE28146	Hippocampus	Braak stage at V to VI	Affymetrix Human Genome U133 Plus 2.0 Array	30	1
	GSE29378	Hippocampus	Braak stage at V to VI	Illumina HumanHT-12 V3.0 expression beadchip	63	2
	GSE32645	Cortex	Braak stage VI	Agilent-014850 Whole Human Genome Microarray 4 × 44K G4112F	6	0
	GSE33000	Prefrontal cortex	–	Rosetta/Merck Human 44k 1.1 microarray	467	30
	GSE36980	Frontal cortex, hippocampus, temporal cortex	Braak stage at V to VI	Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]	79	8
	GSE37263	Neocortex	Braak stage at III to VI	Affymetrix Human Exon 1.0 ST Array [transcript (gene) version]	16	1
	GSE39420	Posterior cingulate	Braak stage at V to VI	Affymetrix Human Gene 1.1 ST Array [transcript (gene) version]	21	0
	GSE44768	Cerebellum	–	Rosetta/Merck Human 44k 1.1 microarray	230	10
	GSE44770	Prefrontal cortex	–	Rosetta/Merck Human 44k 1.1 microarray	230	28
	GSE44771	Visual cortex	–	Rosetta/Merck Human 44k 1.1 microarray	230	22
	GSE48350	Entorhinal cortex, hippocampus, post-central gyrus, superior frontal gyrus	Braak stage at II to VI	Affymetrix Human Genome U133 Plus 2.0 Array	253	16
	GSE5281	Entorhinal cortex, hippocampus, medial temporal gyrus, posterior cingulate, primary visual cortex, superior frontal gyrus	–	Affymetrix Human Genome U133 Plus 2.0 Array	161	11
	GSE61196	Choroid plexus	Braak stage at III and VI	Agilent-014850 Whole Human Genome Microarray 4 × 44K G4112F	21	1
	GSE84422	Amygdala, anterior cingulate, caudate nucleus, dorsolateral prefrontal cortex, frontal pole, hippocampus, inferior frontal gyrus, inferior temporal gyrus, middle temporal gyrus, nucleus accumbens, occipital visual cortex, parahippocampal gyrus, posterior cingulate cortex, precentral gyrus, prefrontal cortex, putamen, temporal polesuperior parietal lobule, superior temporal gyrus	Braak stage at I to VI	Affymetrix Human Genome U133A ArrayAffymetrix Human Genome U133B ArrayAffymetrix Human Genome U133 Plus 2.0 Array	1,146	64
	GSE93885	Olfactory bulb	–	Affymetrix Human Gene 2.0 ST Array [transcript (gene) version]	18	0
RNA-Seq	GSE104704	Lateral temporal lobe	Braak stage at V to VI	Illumina HiSeq 2500 (Homo sapiens)	30	–
	GSE95587	Fusiform gyrus	Braak stage at III to VI	Illumina HiSeq 2500 (Homo sapiens)	117	–
	GSE53697	Dorsolateral prefrontal cortex	Braak stage at II to VI	Illumina HiSeq 2500 (Homo sapiens)	17	–
	GSE67333	Hippocampus	Braak stage at V to VI	Illumina HiSeq 2000 (Homo sapiens)	8	–

→ The flow diagram of datasets selection, including identification, screening, eligibility and inclusion stage.



→ Dataset preprocessing and identifying DEG'S

1) After dataset selection, all included datasets contained complete transcriptomic raw data, and were undergone a quality control. In the included microarray datasets, the quality of the normalized dataset was assessed using the R package

arrayQualityMetrics. Different numbers of arrays were identified as outliers and omitted in the subsequent analysis.

2) After evaluating the sequence quality, the reads in the included RNA-Seq datasets were mapped to the Homo sapiens reference genome (GRCh38.94). The mapping rates were above 95% for the RNA-Seq datasets. From the mapped reads, the genes were counted to determine DEGs.
3) After preprocessing steps, different numbers of DEGs were determined from each comparison, ranging from 0 to over 10,000.There were 21,064 DEGs in all comparisons that compared AD brain samples with healthy brain samples, and 16,810 DEGs (79.76%) were reported in at least two comparison.
4) In the meta-analysis of DEGs, 9,298 DEGs were found to be statistically significant , 4,960 genes were downregulated and 4,338 genes were upregulated. The most reported downregulated DEGs were DPP6 and FXYD7, which were reported in 16 comparisons; RHOQ was the most reported upregulated DEG across all 15 reported comparisons. A loss of DPP6 or FXYD7 is reported to dysregulate neuronal excitation (Hoos et al., 2013; Cacace et al., 2019), while RHOQ is reported to enhance Aβ oligomerization (Aguilar et al., 2017).
5) **In** the enrichment analysis, AD itself was identified as significant. In the AD pathway, components related to the mitochondrial respiratory chain were downregulated, while components related to Ca2+ **channels, which** transport Ca2+ from the ER **to** the **cytoplasm,** were upregulated. Several AD-related pathways, including **the** proteasome, oxidative phosphorylation, and retrograde endocannabinoid signaling **pathway,** were also identified as statistically significant; most components **of** these pathways were downregulated.
6) **Downregulation** of the proteasome **reduces Aβ** clearance, while **attenuation** of oxidative phosphorylation **reduces** the **efficiency** of the mitochondrial respiratory chain, suggesting hypometabolism in the AD brain**.**

Quality Assessment of studies

- MIAME and MINSEQE guidelines were used to assess the transcriptomic analysis of studies that published the datasets that were used in the present study. The results of quality assessments are shown in Figure 6. Among the 19 studies that published microarray datasets, 80% did not provide information about the brain sample size or weight used for RNA extraction, and only two studies

provided sufficient information about experimental design, including quality control of samples.

Figure 6

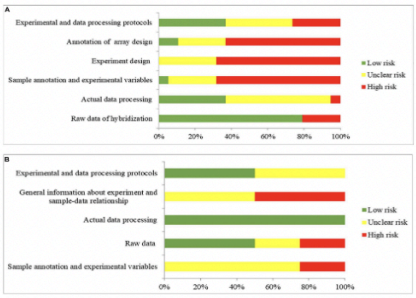


FIGURE 6. The overall quality assessment for **(A)** 19 studies for DNA microarray and **(B)** 4 studies for RNA-seq. Green color represents low risk of bias, which the authors clearly provided the information with full detail. Yellow color represents unclear risk of bias, which the authors provided the information without full detail. Red color represents high risk of bias, which the authors did not provide the correct information.

- The overall quality assessment of 19 studies for DNA microarray and 4 studies for RNA-seq is low, with full detail provided by the authors.

- The quality of the studies was low, and the results might be influenced, blurring some pathological events in the postmortem AD brains. The genes encoded heat shock proteins and neuroinflammation were not found statistically significant in the meta-analysis.