Bioinformatic Profiling of Candidate Gene Involved in Endometrial Cancer Development and Progression.

Li-Yun CHUEH

Abstract

Endometrial Cancer (EC) is a common type of cancer developed in the uterus. EC is the most common type of cancer in uterine cancers for its high incidence and mortality rate. There is a lack of precise biomarkers for early diagnosis. This study aims to identify the potential crucial genes to provide insight into a biomarker in EC cancer progression by bioinformatics analysis.

GSE17025 and GSE39099 were downloaded from Gene Expression Omnibus (GEO) for EC gene profiling. The gene expression pattern of GSE17025 was stage I EC. 584 differentially expressed genes (DEGs) were screened with the control comparison. Following with two strategies to identify functional genes, (1) Protein-Protein Interaction (PPI) network construction with hub gene clustering and (2) gene co-expression network construction from case-control associated modules (WGCNA). The overlapping genes between those two strategies were selected as candidate genes. Further, GSE39099 was used to validate specific target genes expressed in the early stages. The gene expression pattern was from different stages of EC (1) atypical endometrial hyperplasia, (2) stages I and II EC, (3) stages III and IV EC, and the control. 440 DEGs were acquired and performed to the same method as a strategy for hub gene clusters. With the validation hub gene clusters, the target gene of the overlapping was *Matrix Metalloproteinase-2* (*MMP2*).

MMP2 was calculated as a down-regulated gene in both GSE17025 and GSE39099 datasets. Kaplan-Meier survival rate in *MMP2* demonstrated that low *MMP2* was associated with shorter survival (p= 0.005) compared with high *MMP2* expression in uterine corpus endometrial carcinoma patients (n=543). Those findings provided that *MMP2* might be the potential target in EC progression. Furthermore, the pipeline is suitable for functional gene identification discovery.

Index Terms

Endometrial cancer (EC), differentially expressed gene (DEG), protein-protein interaction (PPI) network, co-expression network, *Matrix Metalloproteinase-2* (*MMP2*)

I. INTRODUCTION

Cancer of the endometrium is the most common gynecologic disease affecting the female reproductive tract [1]. In Taiwan, EC is the most common type of cancer in uterine cancers for its high incidence and mortality rate, and its incidence is increasing [2]. EC usually occurs in postmenopausal women; however, there is a rising tide of early onset. Hormone imbalance and obesity increase the risk of developing EC, which is associated with complex gene expression profiles in the progression of EC. Moreover, the staging of EC allowed the doctor to plan the best treatment and to predict a patient's prognosis.

With bioinformatics analysis, the gene expression variable caused by abnormally functioning genes [3] gives us a hint to screen out the core genes for the assessment of diagnostic value. In this study, the DEGs of stage I EC was screened from GSE17025 and subjected to PPI or gene co-expression network constructions (WGCNA) to select the overlapping genes as candidates. Furthermore, the DEGs in the different stages of EC in GSE39099 was used to validate specific target genes expressed in the early stages.

II. METHODS

Data collection

The gene expression profiling was downloaded from Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/). The microarray platforms for GSE17025 and GSE39099 were Platform GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. Dataset GSE17025 involved in 91 cases of stage I EC tumors and 12 cases of atrophic endometrium in postmenopausal women as control. Dataset GSE39099 was 4 groups of the pooling samples, involved in normal endometrium as control, atypical endometrial hyperplasia, early-staged (stages I and II) tumors, and advanced-staged (stages III and IV) tumors.

Pipeline in this study

Flow chart depicting the steps in this study pipeline (**Fig. 1**) to distinguish functional genes in endometrial cancer progression.

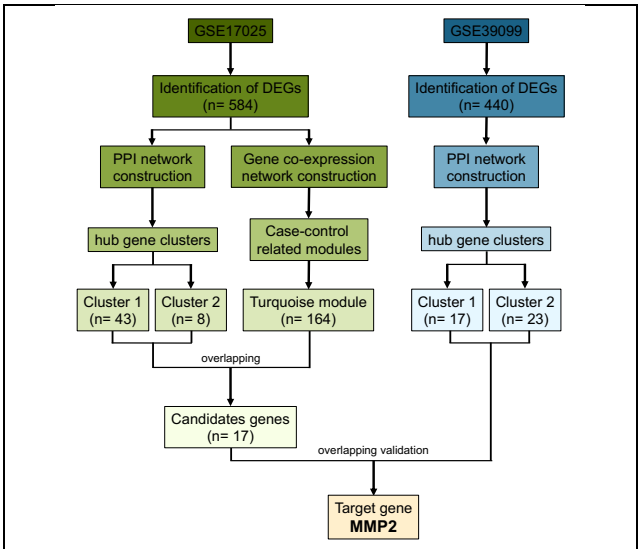

Fig 1. Pipeline in this study for exploring functional gene in endometrial cancer progression.

Data preprocessing and differentially expressed genes (DEGs)
Pre-processing of raw data followed with log2 transformation and confirmation of sample quality by observing expression distribution between all samples. In GSE17025, DEGs screened the control comparison using "limma" R language package. As a set of contrast matrix for calculation with empirical Bayes statistics for gene set testing to fit the linear model. In GSE39099, DEGs were screened using "edgeR" R language package. Assuming control, not DEGs, for dispersion value estimation to fit the generalized linear model. The criteria were adjusted $p<0.05$ and the value of $|\log2\text{fold change}| > \text{mean}(|\log2\text{fold change}|) + 2\times\text{sd}(|\log2\text{fold change}|)$ (the mean value plus 2 times of standard deviation of value).

Comprehensive analysis of differentially expressed genes (DEGs) Protein-Protein interaction (PPI) network construction and hub genes clustering
The DEGs from two datasets were subjected to construct PPI network assessed by Search Tool for the Retrieval of Interacting Genes/Proteins (STRING: https://www.string-db.org/) database [4] and visualized network with Cytoscape [5]. Furthermore, Molecular Complex Detection (MCODE), the PPI network clustering algorithm, was used to select hub gene clusters. The criterion was set as degree cut-off=2, node score cut- off=0.2, k-core=2, and max. Depth=100 according to previous instruction [6].

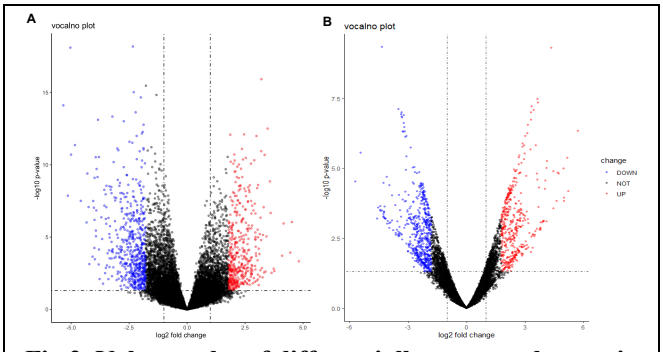

Fig 2. Volcano plot of differentially expressed genes in (A) GSE17025 and (B) GSE39099 dataset.

Gene co-expression network construction
The co-expression network constructed from the DEGs in GSE17025. The "WGCNA" R language package, was used to compose Pearson's correlation matrices to extract the eigenvalues for module eigengenes [7-9]. The soft-thresholding were set according to previous instruction. The network construction and module detection were automatically established based on the soft-thresholding power. Consequently, the cluster dendrogram and the correlation heatmap of module eigengenes could be generated. The case-control matrix (the contrast matrix) was used to clarify the correlation between modules and cancer development.

The correlation between *MMP2* expression and survival
Kaplan-Meier survival plot was generated by KM plotter (https://kmplot.com/analysis/) [10]. *MMP2* expression levels in the uterine corpus endometrial carcinoma patients (n=543) were from GEO, EGA, and TCGA databases.

III. RESULTS

DEGs in endometrial cancer
The total number of genes was 20824. In GSE17025, the DEG analysis by "limma" package and criteria filtering of adjusted $p<0.05$ and the log2fold change cutoff (as mentioned in methods). In **Fig. 2a**, 584 DEGs for 236 up-regulated and 348 down-regulated genes were identified. In GSE39099, the DEG analysis by edgeR package to compute generalized linear model for assuming dispersion value. In **Fig. 2b**, there were 440 DEGs for 231 up-regulated and 209 down-regulated genes identified.

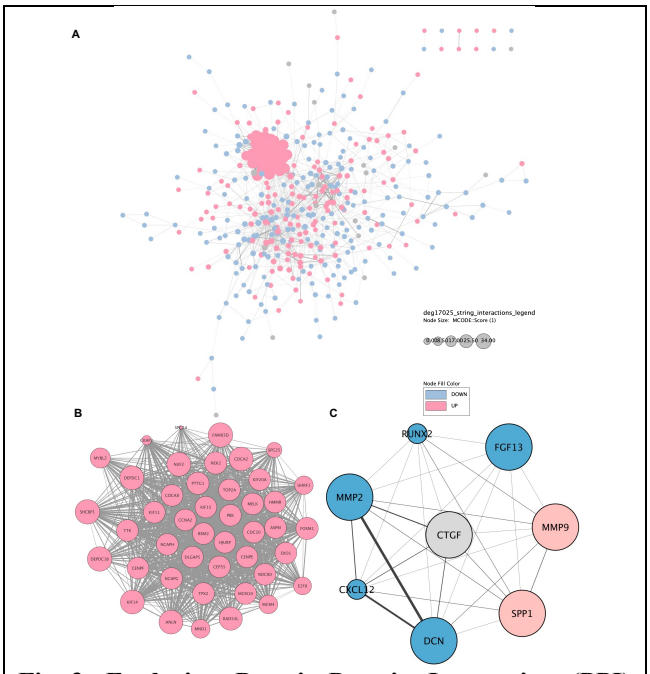

Fig 3. Exploring Protein-Protein Interaction (PPI) networks and clusters of differentially expressed genes (DEGs) in GSE17025.
(A) STRING-predicted [5] PPI network following with visualizing the relationship by Cytoscape [11]. (B-C) Molecular Complex Detection (MCODE) clustering hub genes. (B) cluster 1 included 43 nodes and 860 edges and (C) cluster 2 included 8 nodes and 28 edges.

Exploring Protein-Protein Interaction networks and clusters of differentially expressed genes in GSE17025.
The 584 DEGs were next subjected to STRING database [4], with 509 nodes and 1947 edges, including physical and functional associations. In **Fig. 3a**, the visualization of the relationship by Cytoscape [11], the pink nodes were upregulated genes, the blue was downregulated, and the grey was non-differentially expressed. The edge width corresponded to the co-expression level. Next, we used Molecular Complex Detection (MCODE) to select hub gene clusters (criterion mentioned in methods). The node size of the PPI network corresponded to the MCODE score. For each node with at least 2 edges (k-core= 2), 11 clusters were calculated. With the clustering results, cluster 1 (**Fig. 3b**) comprised 43 nodes and 860 edges, with the highest score, 40.952. Cluster 2 (**Fig. 3c**) was composed of 8 nodes and 28 edges with a score of 8. The top 2

hub gene clusters suggested that those candidates' DEGs (cluster 1 and 2 for 51 genes) might participate in EC progression.

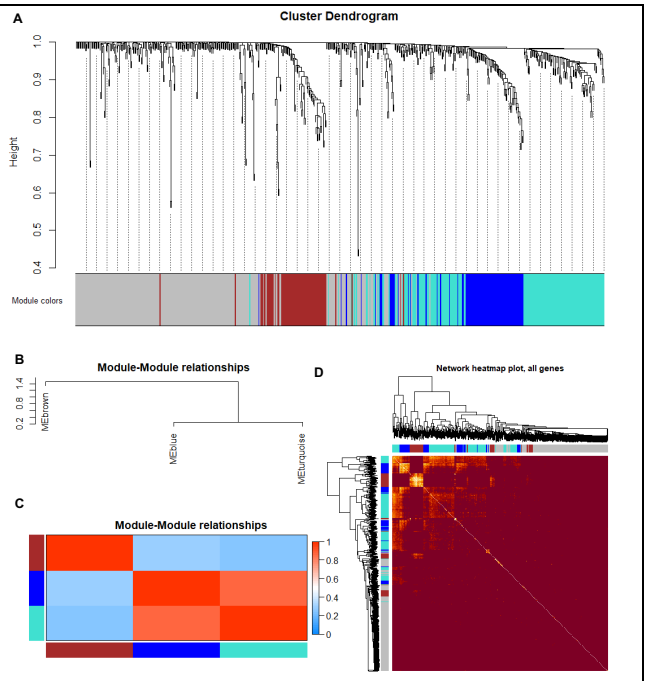

Fig 4. Differential co-expressed gene networks and modules.
(A) Average-linkage hierarchical clustering dendrogram of the DEGs. Module eigengene adjacency (B) dendrogram and (C) heatmap. (D) Heatmap of Topological Overlap Matrix (TOM) among all genes.

The relationship between differential co-expressed gene networks and EC-related modules by WGCNA.

For finding the clusters of highly correlated genes, the "WGCNA" package in R language was used for weighted gene co-expression. With the soft thresholding power matching scale-free topology, the power is set at 6, and 3-color of modules were constructed automatically (**Fig. 4a**). Among them, Pearson's correlation of these 3 co-expression modules' dendrogram (**Fig. 4b**) and correlation heatmap (**Fig. 4c**) showed that module turquoise and module blue were the same patterns and module brown was a separate cluster at first. **Fig. 4d** showed the heatmap of all genes within different modules. Topological Overlap Matrix (TOM) was identified as darker color for higher overlapping and vice versa. The relation between the case-control contrast matrix and modules showed that module turquoise was within the highest correlation (Pearson r= 0.91 and p-value= 8×10^{-231}). There were 578 genes in module turquoise, and gene significance and module membership were highly correlated (**Fig. 5**).

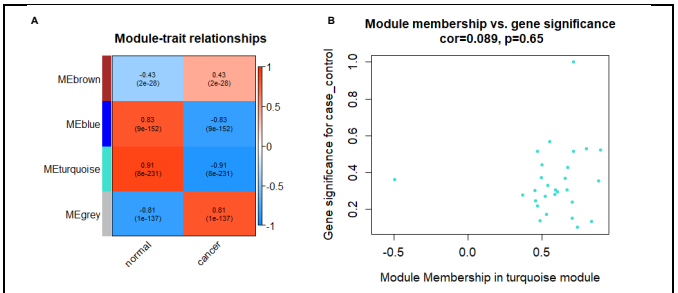

Fig 5. The relationship between endometrial cancer and modules.
(A) The relationship between case-control contrast matrix and modules. (B) The correlation between gene significance and module turquoise membership.

Exploring Protein-Protein Interaction (PPI) networks and clusters of differentially expressed genes (DEGs) in GSE39099.

The 440 DEGs were next subjected to STRING database [4], with 426 nodes and 1164 edges, including physical and functional associations. In **Fig 6a.**, the visualization of the relationship by Cytoscape [11], the pink nodes were upregulated genes, the blue was downregulated, and the grey was non-differentially expressed. The edge width corresponded to the co-expression level. Next, we used Molecular Complex Detection (MCODE) to select hub gene clusters (criterion mentioned in methods). The node size of the PPI network corresponded to the MCODE score. For each node with at least 2 edges (k-core= 2), 9 clusters were calculated. With the clustering results, cluster 1 (**Fig 6b.**) was composed of 17 nodes and 63 edges, with the highest score of 7.875. Cluster 2 (**Fig 3c.**) was composed of 22 nodes and 68 edges with a score of 6.476. The top 2 hub gene clusters suggested that those candidates' DEGs (cluster 1 and 2 for 51 genes) might participate in EC progression.

Kaplan-Meier survival plot for *MMP2* expression with uterine corpus carcinoma.

In DEGs from GSE17025, the overlapping genes are those from the PPI network or gene co-expression network as 17 candidate genes. And further, with the validation by GSE39099, the PPI network of hub gene clusters, the overlapping one, only contains *MMP2*. Additionally, to confirm the possible function of *MMP2* in EC, the Kaplan-Meier survival plot for *MMP2* expression in uterine corpus endometrial carcinoma patients was applied. In **Fig. 7**, the patients with low expression had a low probability of surviving (logrank P= 0.0055).

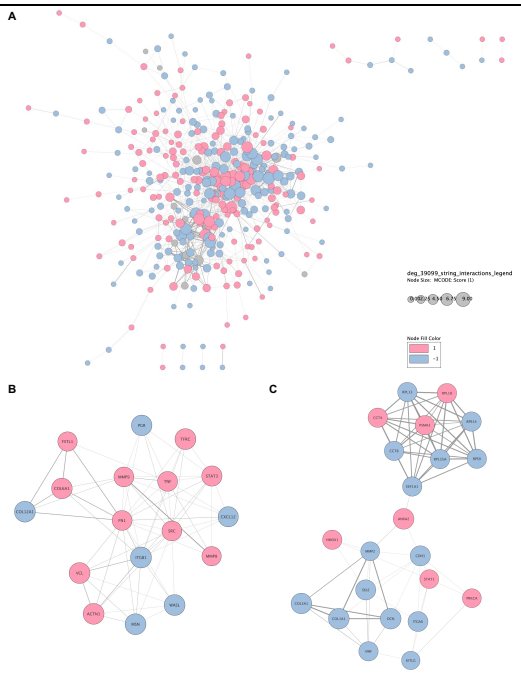

Fig 6. Exploring Protein-Protein Interaction (PPI) networks and clusters of differentially expressed genes (DEGs) in GSE39099.
(A) STRING-predicted [5] PPI network following with visualizing the relationship by Cytoscape [11]. (B-C) Molecular Complex Detection (MCODE) clustering hub genes. (B) cluster 1 included 17 nodes and 63 edges and (C) cluster 2 included 22 nodes and 68 edges.

IV. DISCUSSION

Endometrial cancer is the most common type of uterine cancer. Diagnostic procedures for EC included pelvic ultrasonography with an endometrial biopsy for endometrial thickness measurement, hysteroscopy, blind dilation, and curettage for uncertainty in pelvic ultrasonography [12-15]. Furthermore, Computed tomography or Magnetic Resonance Imaging was used to assess the invasion or metastasis of pelvic lymph nodes. With the costly, invasive, and time-consuming diagnostic process, it is essential to research a precise genetic biomarker for early EC development and progression diagnosis. In this study, we used 2 datasets, GSE17025 and GSE39099, from Gene Expression Omnibus to examine EC gene profiling. GSE17025 included 12 normal and 91 tumor tissue samples with stage I EC. At first, the differentially expressed genes were filtered with adjusted $p<0.05$ and log2fold change at the cutoff of $|\log2\text{fold change}| > \text{mean}(|\log2\text{fold change}|) + 2\times sd(|\log2\text{fold change}|)$, and screened out 584 DEGs, for 236 up-regulated and 348 down-regulated genes. Following with Protein-Protein Interaction network and gene co-expression network constructions (WGCNA), those two strategies to select the hub genes clusters. The intersection of hub gene clusters from 2 were 17 candidates. By contrast, the other dataset, GSE39099, applied validation against the candidate genes from GSE17025 to target the specifically expressed gene in early-staged EC, included 4 groups of the pooling samples from different stages of EC (1) atypical endometrial hyperplasia, (2) stages I and II EC, (3) stages III and IV EC, and the control. Passing with the same cutoff as testing group (the results filtered from GSE17025), and acquired 440 DEGs for 231 up-regulated and 209 down-regulated genes. Next subjected the DEGs to PPI networks to construct hub gene clusters. Last, with the intersect of hub gene clusters from different datasets, the only overlapped, target gene was *Matrix Metalloproteinase-2* (*MMP2*).

MMP2 was one of the most studied genes in the cancer domain. The matrix metalloproteinase (*MMP*) gene family participated in the cleaving extracellular matrix (ECM) and involved in signal transduction, so as *MMP2*. *MMP2* involved in various functions including healing wounds, inflammation, angiogenesis, tumor invasion, and metastasis [16]. The activated *MMP2/MMP9* mediated ECM degradation [17] and regulating angiogenesis [18]. However, we found that *MMP2* were downregulated DEGs both GSE17025 (**Fig. 3c**) and GSE39099 (**Fig. 6c**), and lower expression of *MMP2* with low survival probability in Kaplan-Meier survival plot (**Fig. 7**). Interestingly, on the contrary, the C-terminal non-catalytic fragment of *MMP2* (also known as hemopexin-like domain of *MMP2*), *PEX*, enabled anti-angiogenic, anti-tumor properties, and inhibits cell migration [19, 20]. The disruption of angiogenesis by *PEX* gave us a hint that *MMP2* might not only promoted tumor invasion, but oppositely inhibited tumor growth and migration.

In conclusion, the comprehensive analysis of EC progression enlightened that *MMP2* might be a potential biomarker for early-staged EC diagnosis. Furthermore, the pipeline is suitable for functional gene identification discovery.

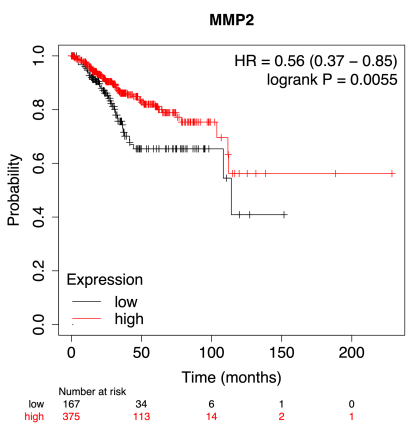

Fig 7. Kaplan-Meier survival plot for *MMP2* expression with uterine corpus carcinoma.

APPENDIX

DEGs of GSE17025										DEGs of GSE39099	
hub gene clusters			co-expression genes							hub gene clusters	
cluster_1		cluster_2	turquoise							cluster_1	cluster_2
PBK	E2F8	CTGF	ABCA17P	CTSF	GBP5	LRRN1	OGN	SEC16B	UBXN4	WASL	VWF
NDC80	NUF2	FGF13	ABRACL	CTSV	GHR	LYN	P2RY14	SEMA5A	UGGT2	STAT3	RPS9
RRM2	ASPM	RUNX2	ACAP2	CXCL10	GLT8D2	LYPD1	P4HB	SERP2	VDAC1	SRC	MMP2
FOXM1	PTTG1	SPP1	ACE	CYS1	GNG11	LYPLAL1	PAK3	SGPP2	VWDE	MMP9	RPL13
HMMR	SHCBP1	MMP2	ADCY6-DT	DGKE	GPSM2	MAMDC2	PAPLN	SHCBP1	WT1-AS	VCL	KITLG
KIF11	TOP2A	DCN	ADHFE1	DOP1A	GSPT2	MAT2A	PAX6	SIAH3	YBX1	ITGB1	RPS14
CCNA2	SPC25	MMP9	ADM	E2F8	HAND2-AS1	MBLAC2	PBK	SIX3	ZBTB17	COL12A1	HMOX1
FAM83D	CENPE	CXCL12	ANGPTL1	ECM2	HAPLN1	MFSD6	PDCD6	SLC39A3	ZFPM2	PGR	SELE
MELK	CDC20		ANLN	EDN3	HEXD	MGC70870	PDSS1B	SLC6A2	ZNF300P1	FN1	STAT1
NCAPG	EXO1		ARID4A	EFEMP2	HJURP	MIDEAS-AS1	PEG3	SLC7A1	ZNF334	FSTL1	ITGAX
CDCA2	CENPF		ARMCX4	EPHX2	HSP90B1	MIER3	PEX1	SLITRK4	ZNF396	MSN	COL3A1
DLGAP5	DEPDC1B		ASPM	EXD2	ID3	MIGA1	PEX3	SNRNP200	ZNF497-AS1	CXCL12	ANXA2
TPX2	TTK		ATF2	FBXL12	IDH2	MIOS-DT	PNISR	SRPRB	ZNF562	TFRC	DCN
CDCA8	HJURP		B2M	FBXO17	KDM2A	MIR100HG	PPIF	SRSF6	ZNF571-AS1	ACTN1	PRKCA
RAD54L	KIF14		B3GALNT2	FBXO4	KIF11	MMP2	PRKAG2-AS1	STK3		COL6A1	CCT4
MND1	MCM4		BEND6	FCRL5	KLF3-AS1	MND1	PSME3	STX16		TNF	EEF1A1
CKAP2	NEK2		CA3	FGD6	KRT6A	MPZL1	PTTG3P	TCEAL2		MMP8	RPL35A
NCAPH	DEPDC1		CCN1	FGF13	KRTDAP	MRTFB	RBM20	TCEAL7			PSMA3
MCM10			CDC20	FOXJ2	LCN2	MYBL2	RBP7	TCF20			CCT8
MYBL2			CENPF	FZD10-AS1	LGALS4	NAA16	ROBO3	TEP1			CDH1
KIF20A			CEP55	GABRP	LOC101927668	NCAPG	RORB	TMEM132C			RPL18
ANLN			CHST6	GALT	LOC105372404	NEBL-AS1	RPL27A	TPX2			COL1A1
UHRF1			CLEC11A	GAS1	LOC105377458	NEK2	RPL5	TRH			
KIF15			CNTN3	GAS2L3	LOC158434	NPL	RPS6KA2	TUBE1			
CEP55			COL4A1	GAS5	LRRK2	NSMCE4A	SDC1	UBE2Q2P13			

REFERENCE

1. Talhouk, A. and J.N. McAlpine, *New classification of endometrial cancers: the development and potential applications of genomic-based classification in research and clinical care.* Gynecol Oncol Res Pract, 2016. **3**: p. 14.
2. *110 年死因統計結果分析.*
3. Narrandes, S. and W. Xu, *Gene Expression Detection Assay for Cancer Clinical Use.* J Cancer, 2018. **9**(13): p. 2249-2265.
4. von Mering, C., et al., *STRING: a database of predicted functional associations between proteins.* Nucleic Acids Res, 2003. **31**(1): p. 258-61.
5. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life.* Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
6. Bader, G.D. and C.W. Hogue, *An automated method for finding molecular complexes in large protein interaction networks.* BMC Bioinformatics, 2003. **4**: p. 2.
7. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
8. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article17.
9. Langfelder, P., et al., *Is my network module preserved and reproducible?* PLoS computational biology, 2011. **7**(1): p. e1001057.
10. Lanczky, A. and B. Gyorffy, *Web-Based Survival Analysis Tool Tailored for Medical Research (KMplot): Development and Implementation.* J Med Internet Res, 2021. **23**(7): p. e27633.
11. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
12. Touboul, C., et al., *Factors predictive of endometrial carcinoma in patients with atypical endometrial hyperplasia on preoperative histology.* Anticancer Res, 2014. **34**(10): p. 5671-6.
13. Group, S.G.O.C.P.E.C.W., et al., *Endometrial cancer: a review and current management strategies: part II.* Gynecol Oncol, 2014. **134**(2): p. 393-402.
14. Lee, D.O., M.H. Jung, and H.Y. Kim, *Prospective comparison of biopsy results from curettage and hysteroscopy in postmenopausal uterine bleeding.* J Obstet Gynaecol Res, 2011. **37**(10): p. 1423-6.
15. Morice, P., et al., *Endometrial cancer.* Lancet, 2016. **387**(10023): p. 1094-1108.
16. Belotti, D., et al., *Matrix Metalloproteinases (MMP9 and MMP2) Induce the Release of Vascular Endothelial Growth Factor (VEGF) by Ovarian Carcinoma Cells: Implications for Ascites Formation1.* Cancer Research, 2003. **63**(17): p. 5224-5229.
17. Wilcock, D.M., et al., *Activation of matrix metalloproteinases following anti-Abeta immunotherapy; implications for microhemorrhage occurrence.* J Neuroinflammation, 2011. **8**: p. 115.
18. Quintero-Fabian, S., et al., *Role of Matrix Metalloproteinases in Angiogenesis and Cancer.* Front Oncol, 2019. **9**: p. 1370.
19. Brooks, P.C., et al., *Disruption of angiogenesis by PEX, a noncatalytic metalloproteinase fragment with integrin binding activity.* Cell, 1998. **92**(3): p. 391-400.
20. Steffensen, B., et al., *Fragmentation of fibronectin by inherent autolytic and matrix metalloproteinase activities.* Matrix Biol, 2011. **30**(1): p. 34-42.

AUTHOR

First Author – Li-Yun CHUEH, daisyccc5959@gmail.com