Quick article summary — Sofwan Nandana

Exploring Core Genes by Comparative Transcriptomics Analysis for Early Diagnosis, Prognosis, and Therapies of Colorectal Cancer

Research Background : Colorectal Cancer (CRC) became a biggest challenge by its high morbidity rate. Early detection rates may increase the survivability of patients. Thus, in this study the researchers investigated the cause of cancer by comparative transcriptomics analysis. The researchers are also investigating early diagnosis, prognosis, and therapies of CRC by bioinformatics tools and databases.

Abstract : The research mainly focuses on analytical and statistical studies for CRC using open access databases. At first, they gathered 252 common Differentially Expressed Genes (cDEG) between control and cancer. The top ten genes become the core genes (CGs) selected by using the GO & KEGG pathway. The CGs revealed regulatory pathway identification, associated with other disease, and functional & pathway. Statistical analysis became supplementary data for target prediction before jumping into molecular docking and MD. The CGs were docked with over 150 CRC-drugs respectively and they selected four best interactions. Molecular dynamic studies played a role in stability of drug performance. Therefore, this study aims for a suitable plan for CRC patients.

Output : Integrated bioinformatic and statistical approach on developing suitable plan for CRC patient

Method:

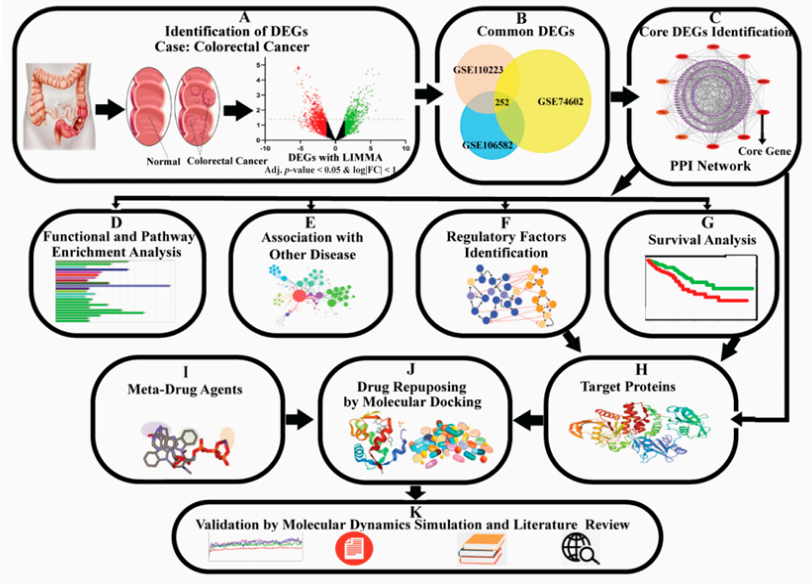

Figure 1. The research plan

a. **Data Mining Microarray & Molecules Set**
   Three microarray datasets CRC caused. CRC IDs GSE106582, GSE110223, GSE74602 were using the platform GPL10558, GLP96, and GLP6104 respectively. The candidate of drugs was obtained from the DSigDB database.
b. **Identification DEGs**

GEO2R web tool based on the LIMMA to identify DEGs between cancer & normal tissue. Benjamini-Hochberg were used to adjust *p*-value. Visualized Venn diagram by FunRich 3.1.3.

c. **PPI Network Analysis**
   CGs were retrieved from the STRING database. Selected from Cytohubba plugin Cytoscape using MCC topology.
d. **Association CGs with CRC Progression**
   Box-plot analysis were performed based on their expression levels in different CRC progression through the UALCAN web tool with the TCGA-COAD & TCGA-READ databases.
e. **Prognosis Power CGs**
   Multivariate Kaplan-Meier were performed using SurvExpress web tool based on TCGA-COAD & TCGA-READ databases.
f. **CGs-Set Enrichment Analysis**
   The Enrichr web tool investigated the association of CGs with terms of interest.
g. **Association CGs with Different Diseases**
   The Enrichr web tool verifies association CGs with other diseases. DisGeNET databases were used to investigate over 21,671 genes and 30,170 disease overlap.
h. **Association CGs with GO Terms and KEGG Pathway**
   The Enrichr web tools were used to obtain Molecular function, Biological process, and Cellular component based on KEGG database.
i. **CGs Regulatory Network Analysis**
   This article compared CGs transcription factor (TFs) vs CGs miRNAs vs CGs through Network Analyst platform-based JASPAR and TarBase databases, respectively.
j. **Molecular Docking & MD simulations**
   They conducted molecular docking to explore FDA-approved drugs for CRC. CGs protein & TFs protein as the target receptor with 14 total.The 3D structures retrieved from Protein Data Bank (PDB) and SWISS-MODEL. PubChem databases were used for 3D molecules with a total 158 drugs. The visualization protein & co-crystal ligand were performed via Discovery Studio Visualizer 2019. The proteins were processed using AutoDock tools & the Swiss PDB viewer adding the structural change & reducing energy receptor, respectively. The docked was analyzed through Discovery Studio Visualizer 2019.

   The best four protein-ligand complexes are selected for next simulations. YASARA software based on the AMBER14 force field was used for MD simulations. At the berendsen thermostat, constant pressure, and a 100 ns MD simulation was examined.

Result:

a. **Identification of DEGs**

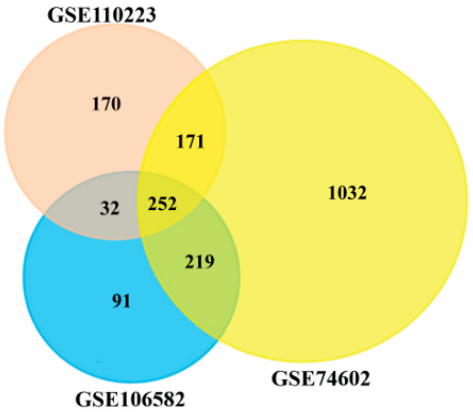

Figure 2. Venn diagram exhibits 252 cDEG from main datasets

b. Identification of Core Genes (CGs)

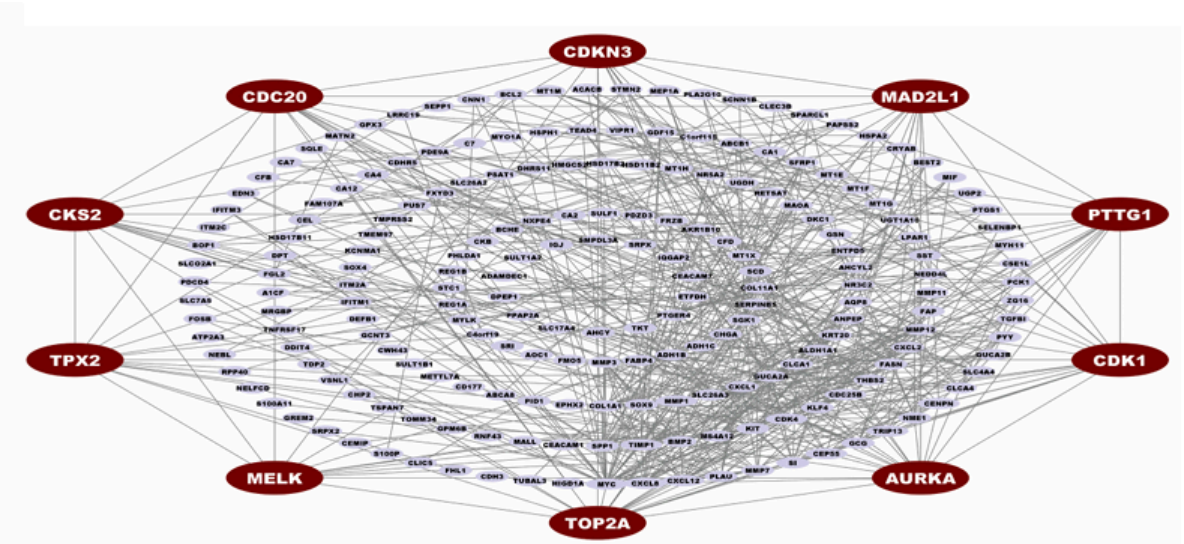

Figure 3. Protein-protein interaction network. The PPI consists of 216 nodes and 616 edges. The red color edges represent CGs.

c. Prognosis power of CGs

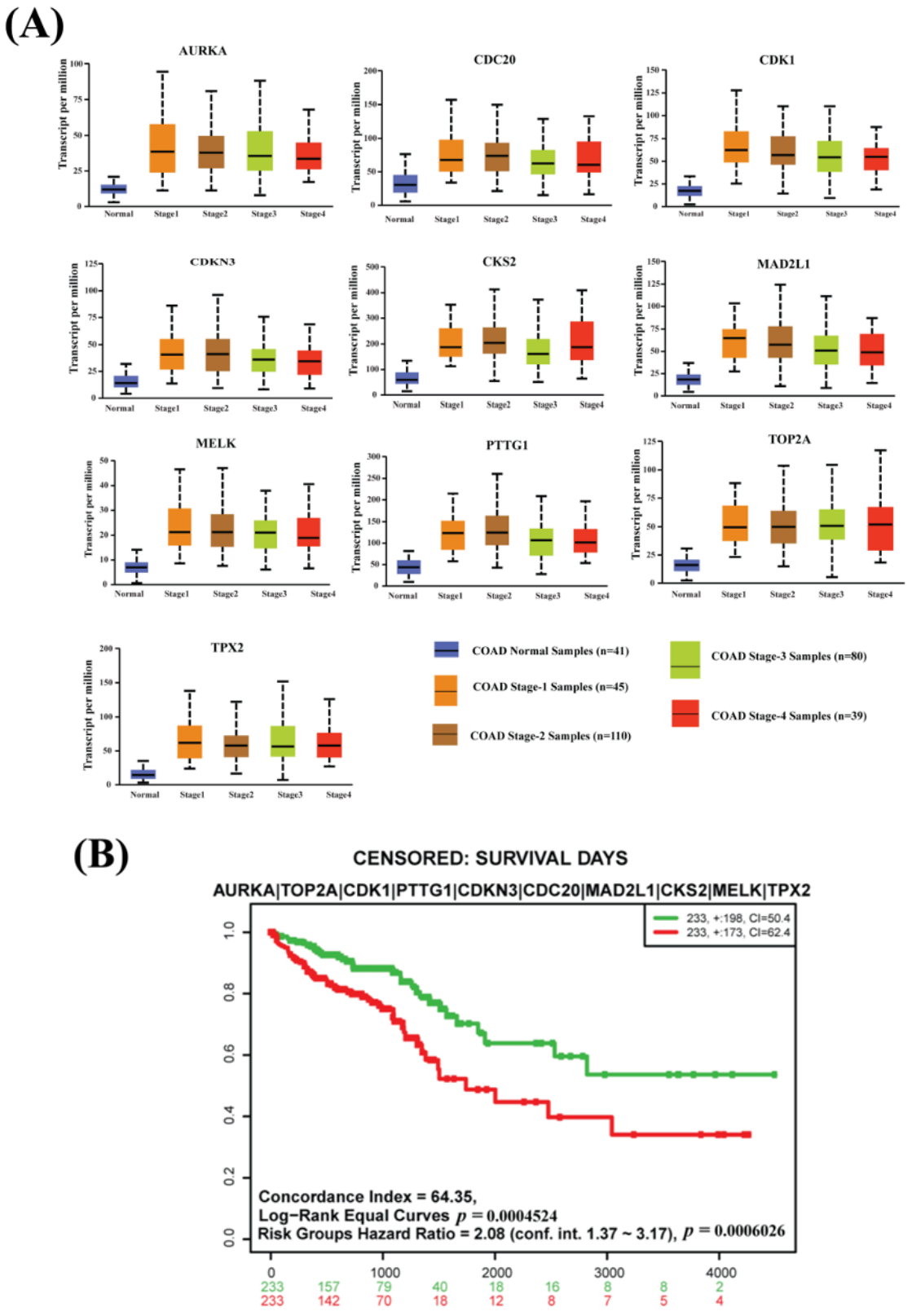

Figure 4. **(A)** Box-plot expressions of CGs with various stages of CRC and normal conditions. **(B)** Multivariate Kaplan-Meier survival probability plot using TCGA-COAD & TCGA-READ databases.

d. Association of CGs with Different Disease

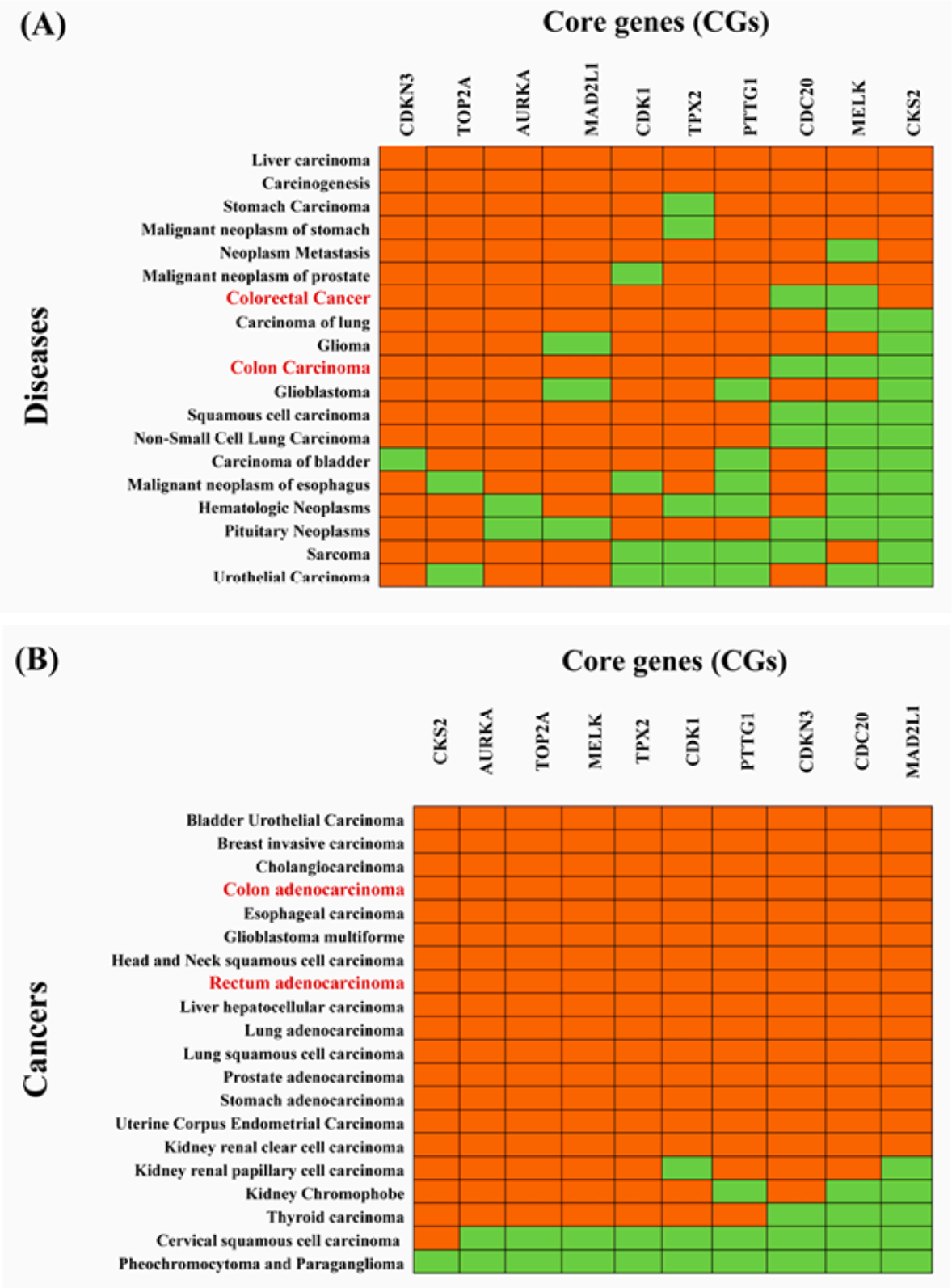

Figure 5. (A) Top 20 Identical CGs disease obtained from Enrich web tool. Orange represents presence; Thus green is identified as absence. **(B)** Top 20 cancer associated with CGs retrieved from pan-cancer analysis based on the TIMER2 web tool with TCGA databases.

e. **Association of CGs with GO Terms and KEGG Pathway**

CGs are involved in many ways with biological processes (BPs), molecular function (MFs), cellular components (CCs), and KEGG pathway. CGs involved in mitotic cell cycle phase transition, anaphase-promoting complex-dependent catabolic process, regulation of G2/M transition of mitotic cell cycle (BPs); histone kinase activity, RNA polymerase II CTD heptapeptide repeat kinase activity, protein kinase binding (MFs); spindle, cyclin-dependent kinase, holoenzyme complex (CCs); and cell cycle, bladder cancer, oocyte meiosis (KEGG pathway).

f. **Identification of Regulatory Factors**

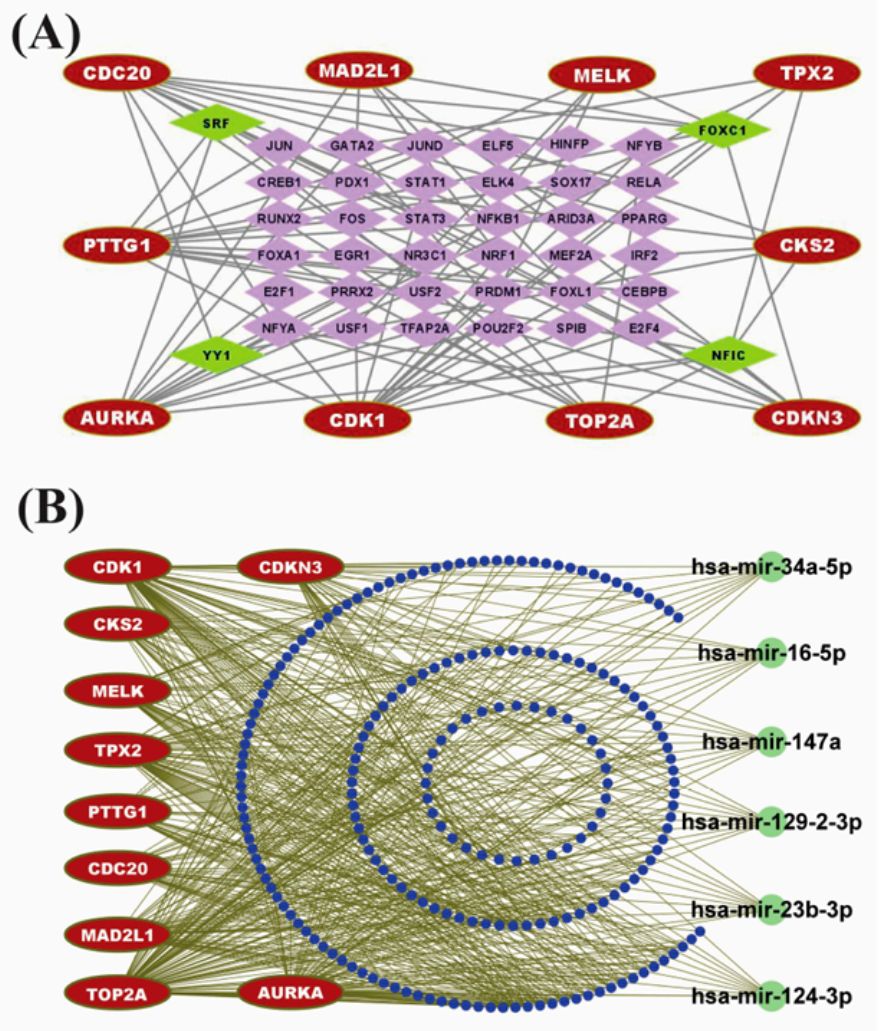

Figure 6. (A) TFs-CGs regulatory network, containing 72 nodes & 174 edges. The red one are CGs, Green & purple represent TFs while green is the core TFs **(B)** Mi-RNA-CGs interaction network, containing 223 nodes & 450 edges. The blue & green represent mi-RNA while the green is the core mi-RNA.

g. **Drug Repurposing through Molecular Docking Studies**

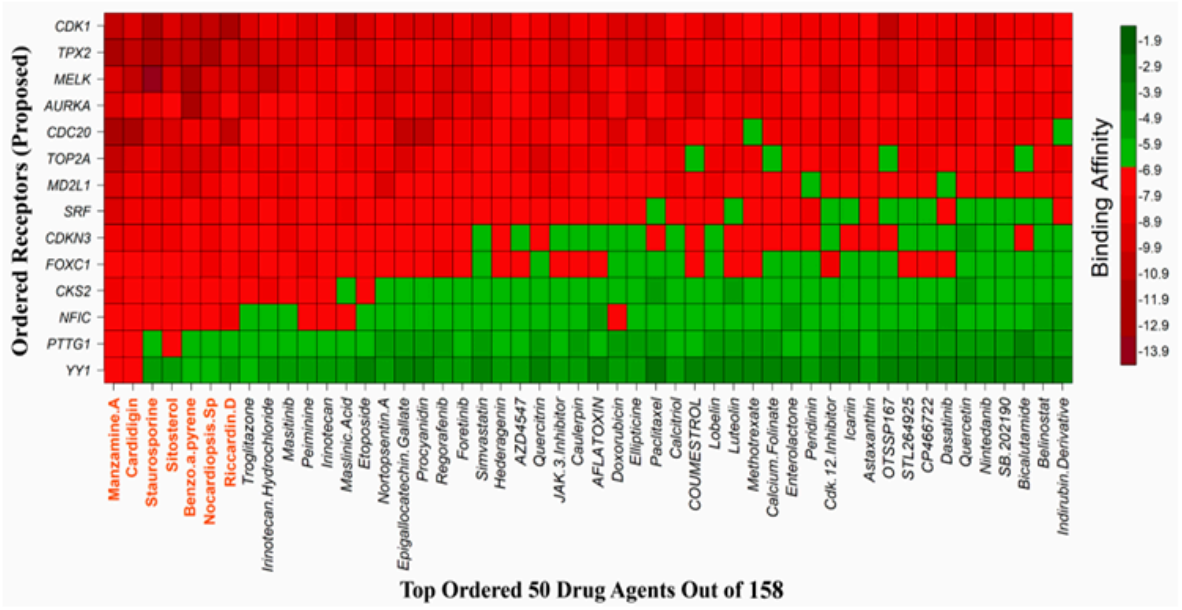

Figure 7. Molecular docking results. The red colors represent strong binding affinity.

h. Molecular Dynamic (MD) Simulations

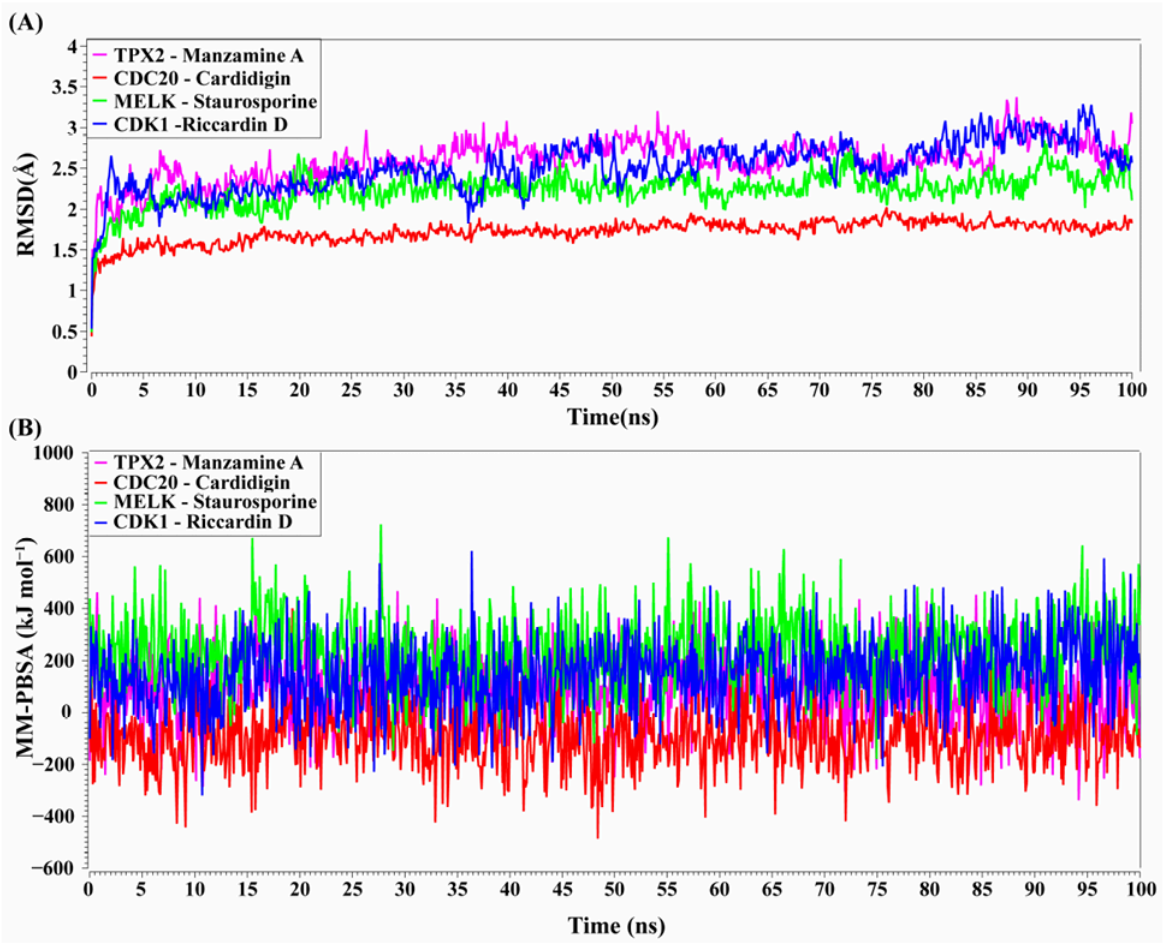

Figure 8. (A) Top-ranked by MD simulations based of RMSDs. **(B)** Top ranked based of binding free energy against each ns simulations.

Discussion: Exploring core genomic activity of CRC may have shortened the research.They identified three datasets and formed the Core Genes (CGs). All the CGs (AURKA, TOP2A, CDK1, PTTG1, CDKN3, CDC20, MAD2L1, CKS2, MELK, and TPX2) show similarity between the genes and CRC. GO terms such as biological process, Molecular functions, cellular component, and KEGG pathway strengthening the evidence. CGs protein were docked with a co-crystal FDA approved drug for seeking the best drug for CRC. Molecular Docking (MD) were performed for further analysis. Based on binding affinity, Manzamine A, Cardidigin, Staurosporine, Benzo[a]pyrene, Sitosterol, Nocardiopsis sp., and Riccardin suggested for CRC patients. In addition, all the data above were validated with several studies.

Conclusion: This study systematically ranks the top 10 core genes (CGs).This study also identified several network studies about CGs TFs and CGs mi-RNA of CRC. Enrichment analysis were performed to reinforce argumentations. In addition, this study also recommends the best seven drugs for CRC therapy plan.

Reference

Islam, M. A., Hossen, M. B., Horaira, M. A., Hossen, M. A., Kibria, M. K., Reza, M. S., Tuly, K. F., Faruqe, M. O., Kabir, F., Mahumud, R. A., & Mollah, M. N. H. (2023). Exploring Core Genes by Comparative Transcriptomics Analysis for Early Diagnosis, Prognosis, and Therapies of Colorectal Cancer. *Cancers*, *15*(5), 1369. https://doi.org/10.3390/cancers15051369.