BIMM 182 Final Report: Project 4

Ella Say, Joseph Hwang, Kai Akamatsu, Kairi Tanaka
esay@ucsd.edu j8hwang@ucsd.edu kakamatsu@ucsd.edu ktanaka@ucsd.edu

Abstract

The goal of our project was to explore the relationship between gene essentiality and ecDNA using the DepMap project. Further, we sought to understand how genetic heritability correlates across tissues and between genes to potentially reveal how proliferation of cancer could be affected by genetic regulation of genes in ecDNA positive cell lines. Gene essentiality can be defined as the impact of knocking out specific genes in various cancer cell lines. Genetic heritability is defined as how much variation in gene expression can be explained by genetics. We wanted to determine if these ecDNA positive cell line critical genes are regulated by genetic variation because we hypothesized that the genetic variants regulating these genes can be genetic risk factors for ecDNA cancers. We aimed to answer two critical questions: "Which genes are essential in cancer cell lines with ecDNA?" and "How are these genes regulated by heritable genetic variations?".

We discovered that the essentiality scores of the genes TMEM 199 and CITED4 are significantly different between ecDNA positive and ecDNA negative cell lines. In our heritability analysis, we used GTEx data to reveal correlations of the estimated heritability between different tissue types. We then classified some genes as more heritable in certain tissues than others. Specifically, for the CITED4 gene of interest, we found the gene was highly heritable in whole blood, thyroid, lung, pancreas, skin, and spleen. The implications of these data range from a greater understanding of the role gene regulation plays in cancer metastasis to biomarkers of early identification.

Methods

To answer the first question, "which genes are essential in ecDNA containing cancer cell lines?", we first had to preprocess and extract the relevant information from our datasets. Using the CRISPR Gene Effect file and aggregated results file, the cell lines were divided into two groups: ecDNA positive and ecDNA negative (**Figure 1**). Next, we identified the subset of genes that were present in both groups. To connect ecDNA presence with gene essentiality scores, we examined whether the essentiality score of a given gene differed significantly between the two groups. A significant difference could provide some indication of which genes were essential for the survival of cancer cells in the presence of ecDNA. As a result, we conducted a two-sample t-test for each gene that was present in both ecDNA and non-ecDNA cell lines. Our null hypothesis was: "For a given gene, there is no difference in the mean essentiality score between ecDNA positive cell lines and ecDNA negative cell lines."After performing the test, we identified 145 genes whose mean essentiality scores differed significantly between ecDNA positive and negative cell lines ($p < 0.05$). Because we performed multiple two-sample t-tests (1,610), there was an increased probability of obtaining false positives. Therefore, we used the Benjamini-Hochberg method with an FDR level of 0.05. After conducting the multiple hypothesis correction, out of the initial 145 genes, we identified two significant genes: TMEM199 and CITED4 (**Figure 2**).

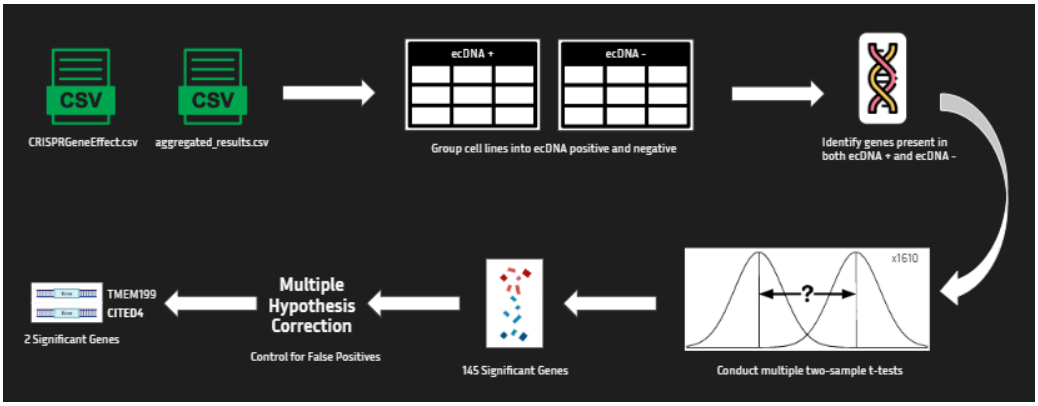

Figure 1. Analysis pipeline starting with comma-separated value files and resulting in two significant genes TMEM199 and CITED4.

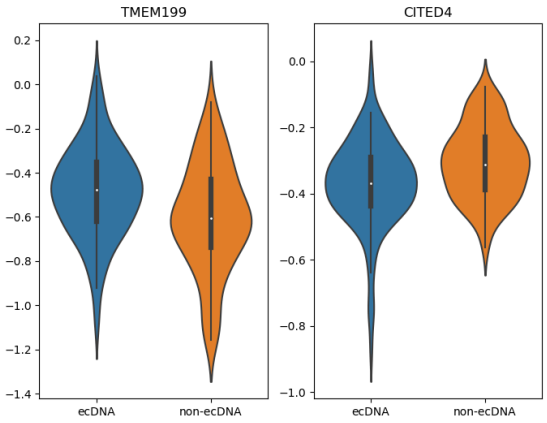

Figure 2. Violin plot comparing the distribution of essentiality scores for TMEM199 (t = 4.07, p = 6e-5) and CITED4 (t = -4.11, p = 5e-5) genes on ecDNA positive and ecDNA negative cancer cell lines.

To begin our analysis of heritability across tissue types, we first extracted genotype and gene expression data across various tissues from the GTEx project. GTEx data, which quantifies the relationship between gene expression and gene variation across multiple different tissues in predominantly European individuals, allowed us to identify the heritability of our 145 significant genes. We applied a tool called GCTA to compute the proportion of variation in gene expression that can be explained by genetics: the heritability of gene expression (Yang et al. 2011 AJHG). GCTA utilizes the genotype and gene expression data to partition the gene expression variance to variance explained by genetics and residual variance. We applied this tool across 14 tissues in GTEx (**Figure 3**). GTEx is based on non-cancerous tissues, but we assume that the heritability values we computed are a good proxy for the heritability values of a cancerous cell from the same tissue. Next, we checked to see if there was a correlation between gene expression heritability across different tissues. A high correlation would indicate that the gene expression heritability of those tissues tends to vary together, while a negative correlation suggested the inverse.

Tissue	Sample Size
Whole Blood	670
Lung	515
Breast	396
Prostate	221

Figure 3. First four tissue types and their sample sizes, analyzed for heritability.

All statistical analyses and visualizations were performed using Python libraries.
Please see our code and summary statistics (FDR_all_genes.csv) on our Github.

Results

Our analysis of tissue correlation showed that the heritability of gene expression is significantly correlated across different tissues. For instance, the heritability of gene expression in colon and stomach tissues was highly correlated ($r = 0.81$). This high correlation suggests that genetic factors influencing gene expression in one tissue type are likely to have a similar influence in another, especially if the tissues are functionally or developmentally related. (**Figure 4**).

Furthermore, some genes were found to be heritable in some tissues and not in others (heritability> 0, $p < 0.01$), emphasizing how gene expression can differ based on tissue type (**Figure 5**). Diving deeper into the genes and tissue heritability, we found that our gene of interest, CITED4, was highly heritable in whole blood, thyroid, lung, pancreas, skin, and spleen (**Figure 6**). This indicates that CITED4's expression is strongly influenced by genetics. The genetic variants regulating the expression of this gene could be genetic risk factors for ecDNA cancers in these tissues.

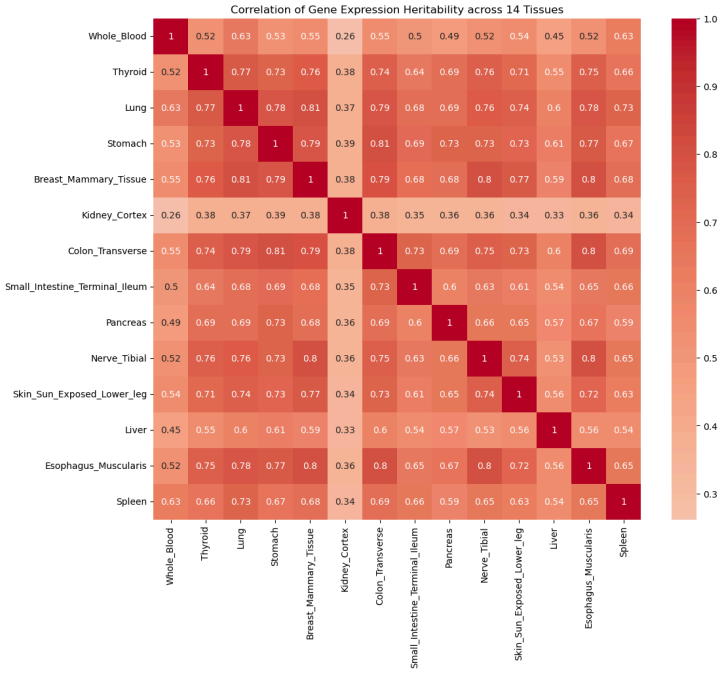

Figure 4. Correlation of gene expression heritability across tissues

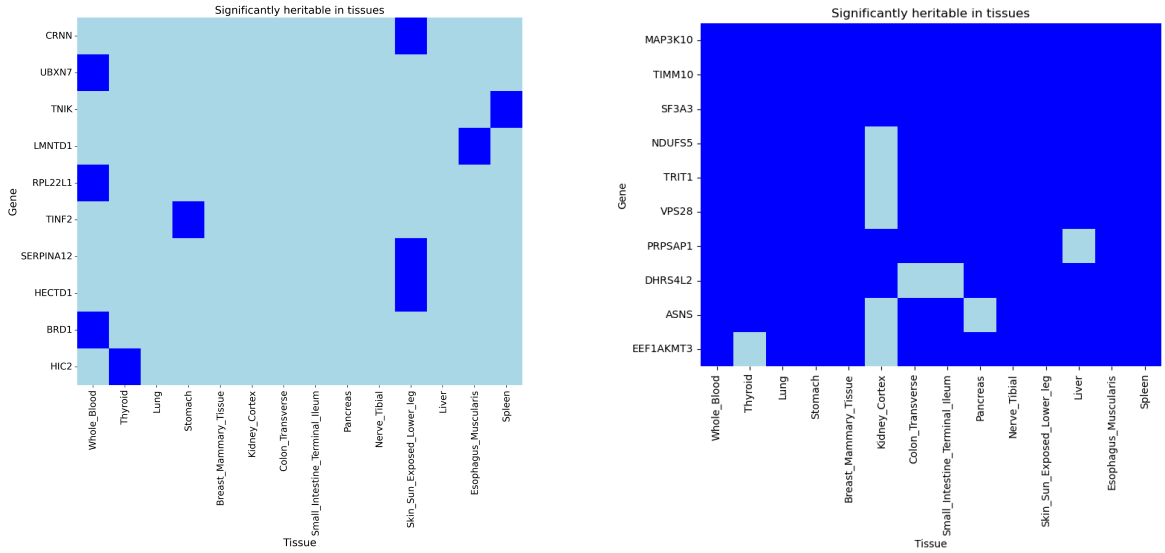

Figure 5. Of the 145 genes for which there was a nominal difference between the essentiality scores of ecDNA + and - cell lines, some genes were significantly heritable in a specific tissue (left) and others in multiple tissues (right)

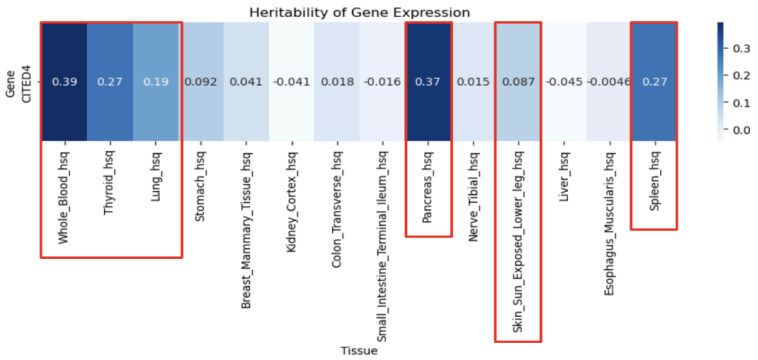

Figure 6. Heritability estimates for CITED4 gene across various tissue types with the significant tissues in red

CITED4 (CBP/p300-Interacting Transactivator with ED-rich tail 4) is a transcriptional coactivator that interacts with CBP/p300, key regulators of gene expression. It plays roles in cellular processes such as differentiation, proliferation, and apoptosis. Previous studies have shown that CITED4 increases metastasis of lung cancer (Zhang et al 2021 Thoracic Cancer). Our heritability analysis shows that the expression of CITED4 in the Lung can be explained by genetic factors, suggesting that the variant regulating the CITED4 expression can be a genetic risk marker for ecDNA positive lung cancer metastasis.

Conclusion

Our study investigated the relationship between essentiality and ecDNA in cancer cell lines, with a specific focus on the heritability of these genes across various tissues. We identified two significant genes, TMEM199 and CITED4, which have significantly different essentiality between ecDNA positive and negative cell lines. Our heritability analysis, conducted using GTEx data, revealed that the CITED4 gene is highly heritable in several tissues, including the whole blood, thyroid, lung, pancreas, skin, and spleen. These findings suggest that genetic regulation plays a critical role in the proliferation and metastasis of ecDNA-driven cancers.

References

NCBI. "CITED4 Cbp/P300 Interacting Transactivator With Glu/Asp Rich Carboxy-Terminal Domain 4 [Homo Sapiens (Human)]." Accessed June 4, 2024. https://www.ncbi.nlm.nih.gov/gene/163732#summary.

Raj, S., Wong, C. T., Tran, T. D., Innes, B. T., Mukherjee, A., Goldthorpe, R. A., & Simpson, K. J. (2023). "Dissecting the Role of EC-DNA in Promoting Tumorigenesis and Therapeutic Resistance." Nature Communications, 14, 39136. https://doi.org/10.1038/s41467-023-39136-7.

DepMap Forum. "Nomenclature of Essential Genes." Accessed June 4, 2024. https://forum.depmap.org/t/nomenclature-of-essential-genes/567.

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2010). "GCTA: A Tool for Genome-Wide Complex Trait Analysis." American Journal of Human Genetics, 88(1), 76-82. https://doi.org/10.1016/j.ajhg.2010.11.011.