BUILDING SEX-SPECIFIC SYNTHETIC LETHALITY NETWORK IN CANCER - TEMP NAME

ALEXANDER TURCO

August 9, 2023

[1] Department of, University, , ON, Canada

Contents

Abstract

Introduction/Background Information

What are Synthetic Lethal Interactions?

Synthetic Lethal Interactions are Harnessed for Precision Oncology

Building Pan-Cancer Synthetic Lethality Networks

Human Sex Differences add An Additional Layer of Complexity

Building Pan-Cancer Synthetic Lethality Networks in a Sex Specific Manner

Materials and Methods

TCGA Data

RNA sequencing (RNA-seq) data was obtained from The Cancer Genome Atlas (TCGA). Raw STAR (Spliced Transcripts Alignment to a Reference) aligned counts for tumor tissue and healthy tissue samples were collected. The Cancer Genome Atlas contains genomic information which spans 33 cancer types. For the purpose of this study, only 12 of the 33 cancer types were considered. We first filtered out sex-biased cancers which include breast invasive carcinoma (BRCA), cervical cell carcinoma (CESC), ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), testicular germ cell tumors (TGCT), uterine corpus endometrial carcinoma (UCEC), and uterine carcinosarcoma (UCS). The reason for this was due to the fact that we are already aware of sex biases in these cancer types. Next, we filtered out blood cancers as well as cancers which lacked normal tissue gene expression samples. This included adrenocortical carcinoma (ACC), lymphoid neoplasm diffuse large b-cell lymphoma (DLBC), glioblastoma multiforme (GBM), acute myeloid leukemia (LAML), and brain lower grade glioma (LGG). The reason for this was due to the fact that we did not have adequate control samples to compare to tumor samples. Finally, we filtered out cancers that did not have any matching pairs of samples (NT and TP from same individual), as well as cancers that had less than 10 matched sample pairs across both males and females. This included mesothelioma (MESO), skin cutaneous melanoma (SKCM), thymoma (THYM), uveal melanoma (UVM), cholangiocarcinoma (CHOL), pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), rectum adenocarcinoma (READ), and sarcoma (SARC). We selected matched tumor-normal sample pairs to help control for genetic background and other individual-specific factors that could influence gene expression. Once cancer types were selected, two pan-cancer raw count gene expression matrices were created, one for males and one for females.

Table 1: List of 12 TCGA Cancer Types With Number of Matched Tumor-Normal Samples in Males and Females.

TCGA information		
Cancer Type	Matched Female Samples	Matched Male Samples
BLCA	9	10
COAD	21	20
ESCA	5	8
HNSC	14	29
KICH	12	13
KIRC	20	52
KIRP	10	22
LIHC	22	28
LUAD	34	24
LUSC	14	37
STAD	10	23
THCA	42	17
TOTAL	213	283

Pre-filtering and Normalization of Raw RNA-seq Count Data

Raw count gene expression matrices for males and females were pre-filtered to remove genes unlikely to exhibit differential expression. For each matrix, we calculated the 90th quantile of overall gene expression as a threshold. For each gene in the matrix, we checked to see whether its expression was greater than the threshold in at least 1 sample. We removed genes where no samples showed an expression value greater than the quantile threshold.

The pre-filtered matrices were then normalized using the DESeq2 package in R (Love et al. 2014). RNA-seq data must be normalized in order to account for factors that prevent the direct comparison of expression measures. The DESeq2 package employs a median of ratios normalization method to account for the inherent biases associated with RNA-seq data. Both sequencing depth (# of reads generated per sample) and RNA composition (differences in composition of RNA molecules in a sample) are factors accounted for by the DESeq2 package. Raw counts are divided by size factors determined for each sample by computing the median ratio of gene counts relative to the geometric gene calculated per gene (Love et al. 2014). Figure 1 summarizes the effects of normalization for both male and female RNA-seq data. The distribution of counts across samples becomes much more consistent, thus making the samples comparable.

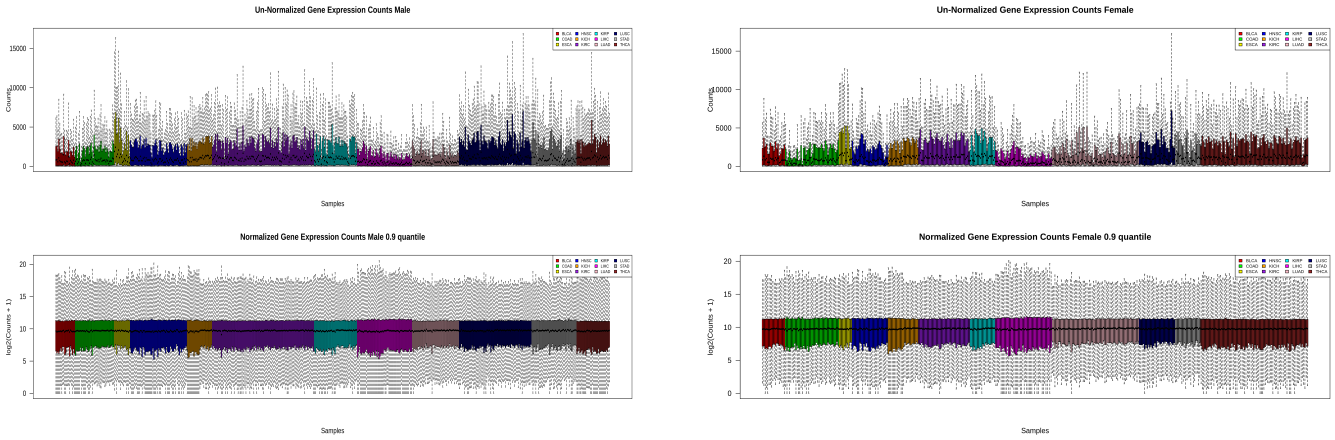

Figure 1: Boxplots highlighting the distribution of raw RNA-seq data (top row) vs normalized RNA-seq data (bottom row) in males (left) and females (right) across normal tissue and tumor tissue samples from 12 TCGA cancer types. Each colour represents a specific cancer type.

Data Quality Assessment (PCA, NPManova)

Once expression matrices were normalized, we performed principle component analysis (PCA) on the normalized counts to gain insights into potential factors contributing to the overall variance. Gene expression data is complex due to large number of variables (genes) present. PCA is a technique used to reduce dimensionality by transforming the data to a new set of variables (principle components) that summarize features of the data (Yeung and Ruzzo 2001). In the context of this study, PCA is useful because it can take expression information from many genes, and reduce it down to fewer dimensions, making the data easier to explore. Using PCA, we explored the effects of tissue condition (normal vs tumor), and cancer type on gene expression.

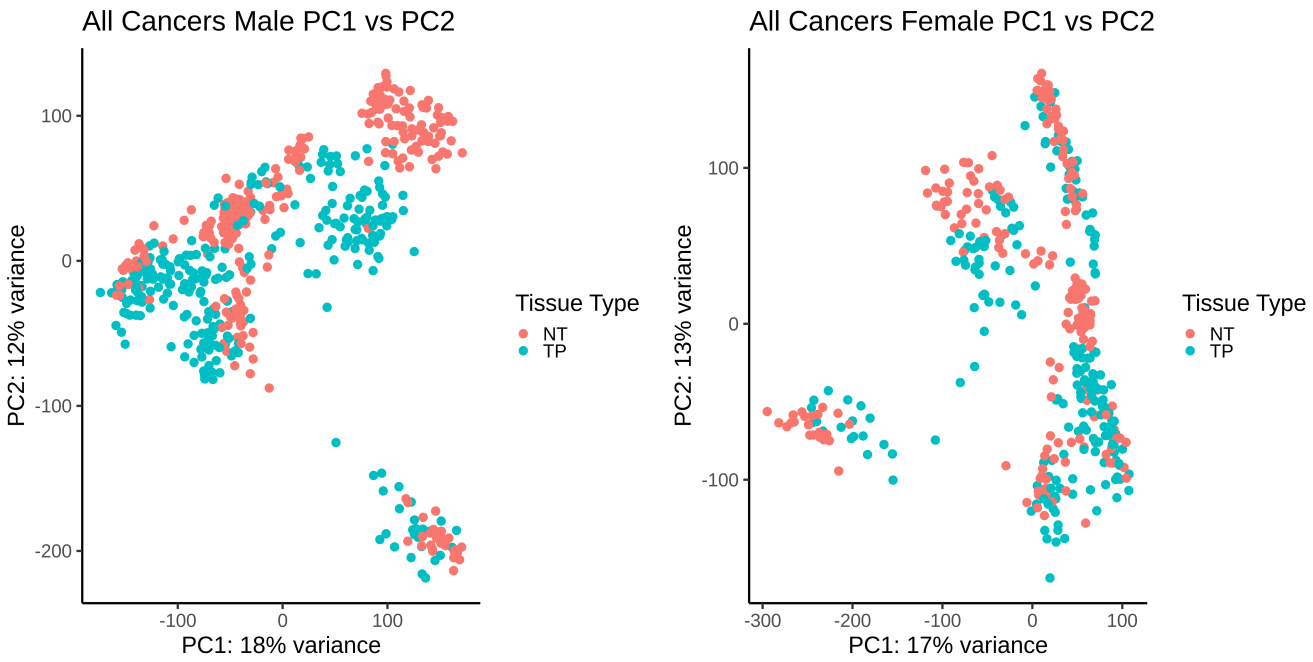

Figure 2: PCA plot highlighting the effect of tissue condition (normal vs tumor) on gene expression in males (left) and females (right) across 12 cancer types. Normal tissue samples are shown in red, tumor tissue samples are shown in blue.

Figure 2 shows the effects of tissue condition (normal vs tumor) on gene expression across the 12 cancer types that were considered in this study. In both males and females, clusters based on tissue condition were present, however there was more variability in clusters of tumor tissue samples compared to clusters of normal tissue samples. Figure 3 shows the effects of both tissue condition and cancer type on gene expression.

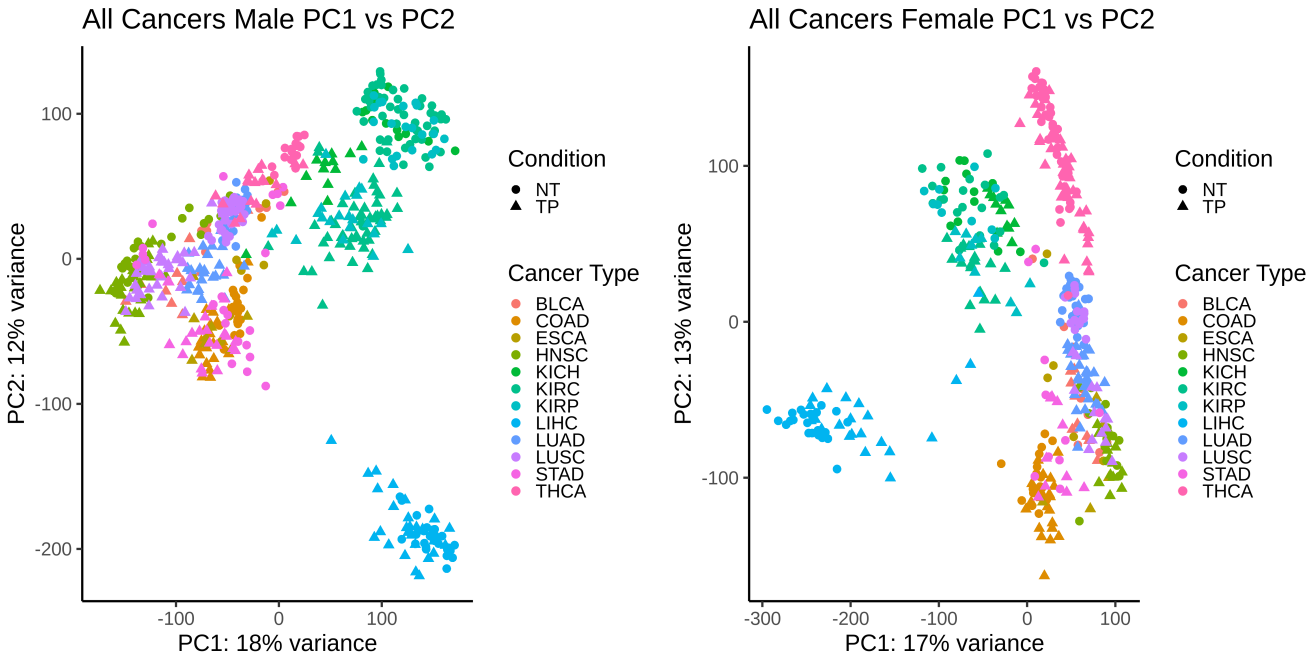

Figure 3: PCA plot highlighting the effect of tissue condition (normal vs tumor) and cancer type on gene expression in males (left) and females (right) across 12 cancer types.

When we added cancer type as a factor, we obtained much clearer clusters in the PCA plot, suggesting that cancer type is a factor that contributes to the variation in gene expression. We wanted to examine factors that contribute to this variance as these factors may influence which genes are found to be differentially expressed.

We also verified the statistical significance of the difference between normal and tumor tissue expression across all cancer types and within each cancer type using a non-parametric multivariate analysis of variance (NPMANOVA), also known as a permutational multivariate analysis of variance (PERMANOVA).

Differential Gene Expression Analysis with DESeq2

Gene Set Enrichment Analysis with ClusterProfiler

Fishers Exact Test For Synthetic Lethality

Next Steps/Future Avenues

Code & Supplementary Materials (Transition Document)

All of the following files listed are located on the Graham Cluster at `~/projects/def-sushant/alextu/slproject_pmcrc`

Scripts used to obtain and format data can be found at `~/projects/def-sushant/alextu/slproject_pmcrc/raw_data/scripts`

Matching executables for these scripts can be found at `~/projects/def-sushant/alextu/slproject_pmcrc/raw_data/executions`

NOTE: THE NUMBERED ORDER THESE SCRIPTS APPEAR IN MATCHES THE ORDER IN WHICH THEY WERE RUN

1. `tcgabiolinks_rawcount_rnaseq_download_female.R` - SCRIPT

   This script utilizes the tcgabiolinks package in R to retrieve gene expression data (RNA-seq) for a specific set of TCGA barcodes. The TCGA barcodes used in this script came from females of a specific cancer type. For each cancer type, a raw gene expression matrix is produced with genes as rows, and samples as columns. This is the first script run in the workflow in order to download the TCGA rna-seq data for the cancer types we were interested in.

   RUN ON LOGIN NODE/LOCAL COMPUTER, NOT ON COMPUTE NODE, COMPUTE NODE CANT DOWNLOAD

2. `tcgabiolinks_rawcount_rnaseq_download_male.R` - SCRIPT

   This script utilizes the tcgabiolinks package in R to retrieve gene expression data (RNA-seq) for a specific set of TCGA barcodes. The TCGA barcodes used in this script came from males of a specific cancer type. For each cancer type, a raw gene expression matrix is produced with genes as rows, and samples as columns. This is the first script run in the workflow in order to download the TCGA rna-seq data for the cancer types we were interested in.

   RUN ON LOGIN NODE/LOCAL COMPUTER, NOT ON COMPUTE NODE, COMPUTE NODE CANT DOWNLOAD

3. `tcgabiolinks_rawcount_clinical_download_female.R` - SCRIPT

   This script utilizes the tcgabiolinks package in R to retrieve clinical metadata for a specific set of TCGA barcodes. The TCGA barcodes used in this script came from females of a specific cancer type. For each cancer type, a SummarizedExperiment (R object) is produced and clinical information such as age at diagnosis and tissue of origin can be extracted.

   RUN ON LOGIN NODE/LOCAL COMPUTER, NOT ON COMPUTE NODE, COMPUTE NODE CANT DOWNLOAD

4. `tcgabiolinks_rawcount_clinical_download_male.R` - SCRIPT

   This script utilizes the tcgabiolinks package in R to retrieve clinical metadata for a specific set of TCGA barcodes. The TCGA barcodes used in this script came from males of a specific cancer type. For each cancer type, a SummarizedExperiment (R object) is produced and clinical information such as age at diagnosis and tissue of origin can be extracted.

   RUN ON LOGIN NODE/LOCAL COMPUTER, NOT ON COMPUTE NODE, COMPUTE NODE CANT DOWNLOAD

5. `extract_clinical_data_females.R` - SCRIPT

   `extract_clinical_data_females.sh` - EXECUTABLE

   This script extracts the clinical information of interest from the SummarizedExperiment (R object). We obtain barcode, tissue, and age information for females in the form of a dataframe.

6. `extract_clinical_data_males.R`- SCRIPT

   `extract_clinical_data_males.sh` - EXECUTABLE

   This script extracts the clinical information of interest from the SummarizedExperiment (R object). We obtain barcode, tissue, and age information for males in the form of a dataframe.

7. `extract_both_tp_nt_samples_females.R` - SCRIPT

   `extract_both_tp_nt_samples_females.sh` - EXECUTABLE

   This script finds all females within a TCGA cancer type who have BOTH normal and tumor samples. The script does not account for individuals with more than one match, meaning that the script will take all tumor and healthy samples pertaining to an individual, not just one pair. In some cases, individuals had two normal tissue samples. The output from this script is a gene expression matrix for each cancer type with only matched normal and tumor samples.

8. `extract_both_tp_nt_samples_males.R` - SCRIPT

   `extract_both_tp_nt_samples_males.sh` - EXECUTABLE

   This script finds all males within a TCGA cancer type who have BOTH normal and tumor samples. The script does not account for individuals with more than one match, meaning that the script will take all tumor and healthy samples pertaining to an individual, not just one pair. In some cases, individuals had two normal tissue samples. The output from this script is a gene expression matrix for each cancer type with only matched normal and tumor samples.

9. `merge_rawcounts_both_tp_nt_samples_females.R` - SCRIPT

   `merge_rawcounts_both_tp_nt_samples_females.sh` - EXECUTABLE

   This script takes the gene expression matrices (with matched Normal-tumor tissue samples) for females of each cancer type and merges them together to create a raw pan-cancer matrix with all 12 cancer types. Rows are genes and columns are samples.

10. `merge_rawcounts_both_tp_nt_samples_males.R` - SCRIPT

    `merge_rawcounts_both_tp_nt_samples_males.R` - EXECUTABLE

    This script takes the gene expression matrices (with matched Normal-tumor tissue samples) for males of each cancer type and merges them together to create a raw pan-cancer matrix with all 12 cancer types. Rows are genes and columns are samples.

    Scripts used to process and visualize data can be found at `~/projects/def-sushant/alextu/slproject_pmcrc/processed_data/scripts`

    Matching executables for these scripts can be found at `~/projects/def-sushant/alextu/slproject_pmcrc/processed_data/executions`

11. `deseq2_extract_normalized_counts_allcancers_female.R` - SCRIPT

    `deseq2_extract_normalized_counts_allcancers_female.sh` - EXECUTABLE

    This script runs the differential gene expression analysis for females (through DESeq2) using the pan-cancer dataframe created from the last script above. This entails normalization of the expression data as well as hypothesis testing for each gene to determine whether expression is significantly different between tumor and normal tissue. This script also gathers all the metadata for the samples included in the pan-cancer matrix (condition (NT vs TP), cancer type, tissue type, age). Lastly, the script can account for different quantile thresholds set by the user. We filtered out genes with expression values less than some quantile value of overall gene expression. We tried 90th quantile and 50th quantile. The output from this script is a DESeq object that can be used to obtain differential gene expression information (raw counts, normalized counts, hypothesis test results for gene expression including log2FoldChange and adjusted p-values).

12. `deseq2_extract_normalized_counts_allcancers_male.R` - SCRIPT

    `deseq2_extract_normalized_counts_allcancers_male.sh` - EXECUTABLE

    This script runs the differential gene expression analysis for males (through DESeq2) using the pan-cancer dataframe created from the last script above. This entails normalization of the expression data as well as hypothesis testing for each gene to determine whether expression is significantly different between tumor and normal tissue. This script also gathers all the metadata for the samples included in the pan-cancer matrix (condition (NT vs TP), cancer type, tissue type, age). Lastly, the script can account for different quantile thresholds set by the user. We filtered out genes with expression values less than some quantile value of overall gene expression. We tried 90th quantile and 50th quantile. The output from this script is a DESeq object that can be used to obtain differential gene expression information (raw counts, normalized counts, hypothesis test results for gene expression including log2FoldChange and adjusted p-values).

13. `deseq2_preprocessing_boxplots_quantile_prefiltering_allcancers_female.R` - SCRIPT

    `deseq2_preprocessing_boxplots_quantile_prefiltering_allcancers_female.sh` - EXECUTABLE

    This script uses the DESeq object produced in the last step for females to extract raw and normalized gene counts for the pan-cancer matrix. The script creates boxplots showing raw counts before normalization vs normalized and transformed

counts. Used to view the effects of normalization. The user can specify the quantile threshold value they would like to use in the script.

14. `deseq2_preprocessing_boxplots_quantile_prefiltering_allcancers_male.R` - SCRIPT
    `deseq2_preprocessing_boxplots_quantile_prefiltering_allcancers_male.sh` - EXECUTABLE

    This script uses the DESeq object produced in the last step for males to extract raw and normalized gene counts for the pan-cancer matrix. The script creates boxplots showing raw counts before normalization vs normalized and transformed counts. Used to view the effects of normalization. The user can specify the quantile threshold value they would like to use in the script.

15. `pca_analysis_mergedcancers_quantiles_prefiltering.R` - SCRIPT
    `pca_analysis_mergedcancers_quantiles_prefiltering.sh` - EXECUTABLE

    This script runs principal component analysis on the normalized count data for males AND females to detect sources of variation in the gene expression data. The user can specify the quantile threshold value they would like to use in the script.

16. `dge_analysis_deseq2.R`

    This script

    RUN ON LOGIN NODE/LOCAL COMPUTER, NOT ON COMPUTE NODE, COMPUTE NODE CANT DOWNLOAD

17. `gsea_analysis_postdeseq.R`

    This script

    RUN ON LOGIN NODE/LOCAL COMPUTER, NOT ON COMPUTE NODE, COMPUTE NODE CANT DOWNLOAD

18. `npmanova.R`

    This script

References

Cheng K, Nair N U, Lee J S, and Ruppin E (2021). Synthetic lethality across normal tissues is strongly associated with cancer risk, onset, and tumor suppressor specificity. Science advances 7(1), eabc2100.

Lee J S, Nair N U, Dinstag G, Chapman L, Chung Y, Wang K, Sinha S, Cha H, Kim D, Schperberg A V, et al. (2021). Synthetic lethality-mediated precision oncology via the tumor transcriptome. Cell 184(9), 2487–2502.

Love M I, Huber W, and Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology 15(12), 1–21.

Shen J P and Ideker T (2018). Synthetic lethal networks for precision oncology: promises and pitfalls. Journal of molecular biology 430(18), 2900–2912.

Shohat S and Shifman S (2022). Gene essentiality in cancer cell lines is modified by the sex chromosomes. Genome Research 32(11-12), 1993–2002.

Yeung K Y and Ruzzo W L (2001). Principal component analysis for clustering gene expression data. Bioinformatics 17(9), 763–774.