Double Trouble: Understanding Sex Differences in Synthetic Lethal interactions in Human Cancers

Project Updates

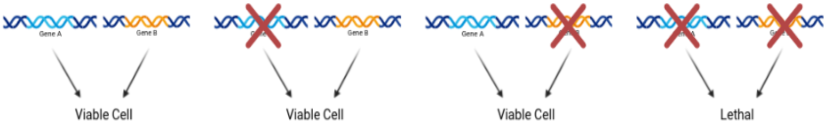
Alexander Turco

July 18, 2023

UHN Princess Margaret Cancer Centre

Recall: The Combined Inactivation of Two Genes Can Lead to Synthetic Lethal Interactions

- Synthetic lethal interactions describe the relationship between two genes whose coupled inactivation, but not their individual inactivation, causes cell death or reduces cell viability



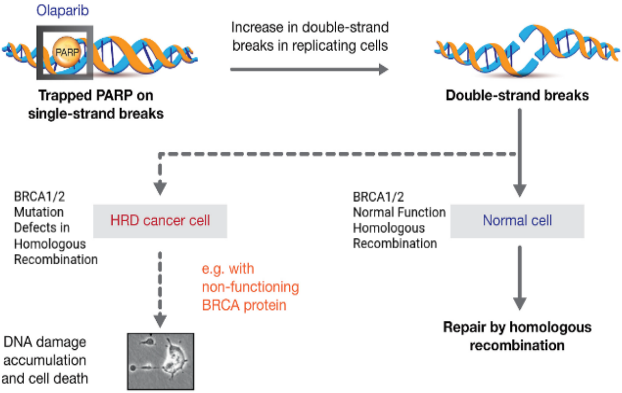
- Inactivation: Preventing or disabling normal function of a gene (e.g mutation)

Synthetic = Combining

[2] Lee et al., 2018

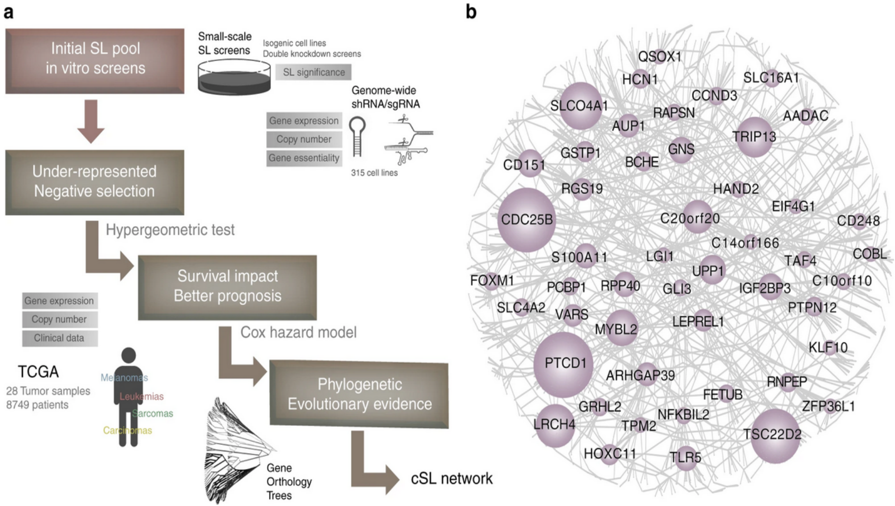
Recall: Synthetic Lethal Interactions are Harnessed for Precision Oncology

- Four FDA approved anti-cancer drugs are Poly [ADP-ribose] polymerase 1/2 (PARP1/2) inhibitors that work via a synthetic lethal mechanism


Olaparib
PARP
Trapped PARP on single-strand breaks
Increase in double-strand breaks in replicating cells
Double-strand breaks
BRCA1/2 Mutation Defects in Homologous Recombination
HRD cancer cell
e.g. with non-functioning BRCA protein
DNA damage accumulation and cell death
BRCA1/2 Normal Function Homologous Recombination
Normal cell
Repair by homologous recombination

[4] Figure From O'Connor, 2015

Recall: Building Pan-Cancer Synthetic Lethality Networks



[2] Lee et al., 2018

Recall: Sex Differences Add an Additional Layer of Complexity

Human sex differences are mainly caused by;

1. Gonadal hormone secretions
2. Genes located on the sex chromosomes (X and Y)

This leads to differences in the frequency of certain cancer types and the efficacy of treatments in males and females

The Objective

Can we build sex-specific synthetic lethality networks for various cancer types?

More specifically, we are trying to elucidate the differences in synthetic lethal interactions between males and females using a network based approach.

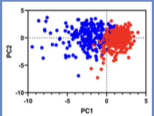
Overall Project Workflow

DATA COLLECTION

	Sample 1	Sample 2
Gene 1	6	3
Gene 2	1438	739
Gene 3	2361	1852
Gene 4	400	951
Gene 5	299	142

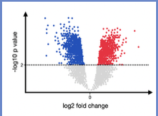
Obtain RNA-seq data from TCGA in the form of raw counts. (Genes are features, samples are data points/variables)

DATA PROCESSING



Normalize and transform raw RNA-seq data, identify sources of variation, batch effects. Do samples cluster according to biological conditions?

DIFFERENTIAL EXPRESSION



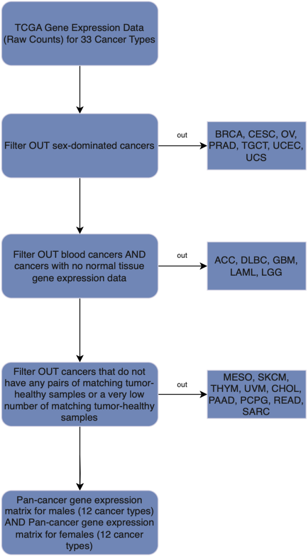
Perform differential gene expression analysis to identify genes that are differentially expressed in tumor tissue.

SYNTHETIC LETHALITY

depmap | Cancer Dependency Map

Find potential candidate SL pairs for differentially expressed genes using CRISPr gene essentiality data from DepMap.

Selection of TCGA Cancer Types



TCGA Gene Expression Data (Raw Counts) for 33 Cancer Types

Filter OUT sex-dominated cancers

out

BRCA, CESC, OV, PRAD, TGCT, UCEC, UCS

Filter OUT blood cancers AND cancers with no normal tissue gene expression data

out

ACC, DLBC, GBM, LAML, LGG

Filter OUT cancers that do not have any pairs of matching tumor-healthy samples or a very low number of matching tumor-healthy samples

out

MESO, SKCM, THYM, UVM, CHOL, PAAD, PCPG, READ, SARC

Pan-cancer gene expression matrix for males (12 cancer types) AND Pan-cancer gene expression matrix for females (12 cancer types)

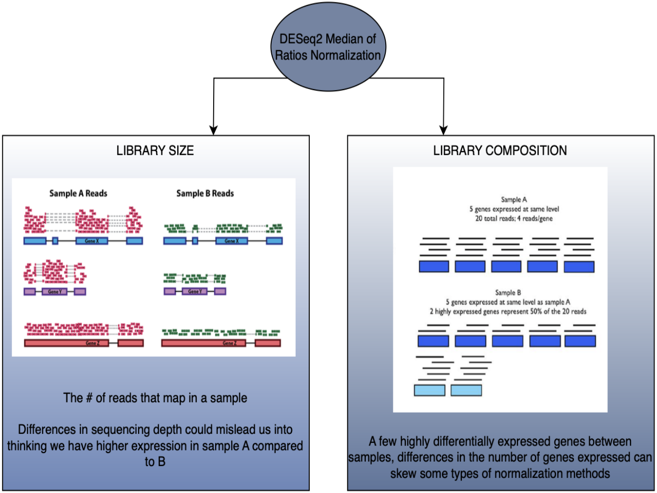
- There is a lack of healthy tissue RNA-seq samples in the TCGA database (12 cancers with more than 10 pairs in M and F)
- Raw count pan-cancer matrices created with 12 remaining TCGA cancer types (BLCA, COAD, ESCA, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, STAD, THCA)

Creating and Pre-Filtering Pan-Cancer Expression Matrices

- For each gene, if expression is $>$ 90th quantile of overall expression in AT LEAST 1 sample, we keep it, otherwise we filter it out
- We reduce the amount of genes with extremely low expression, thereby reducing noise and improving sensitivity to detect differentially expressed genes (genes with weak expression are more susceptible to technical noise arising from library size, library composition, etc)

Normalization for RNA-seq Data Analysis

Required to identify genes that are differentially expressed due to some biological phenomena and not due to technical variation



DESeq2 Median of Ratios Normalization

LIBRARY SIZE

The # of reads that map in a sample

Differences in sequencing depth could mislead us into thinking we have higher expression in sample A compared to B

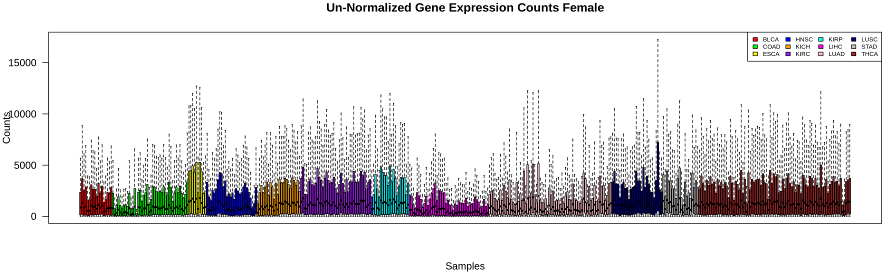
LIBRARY COMPOSITION

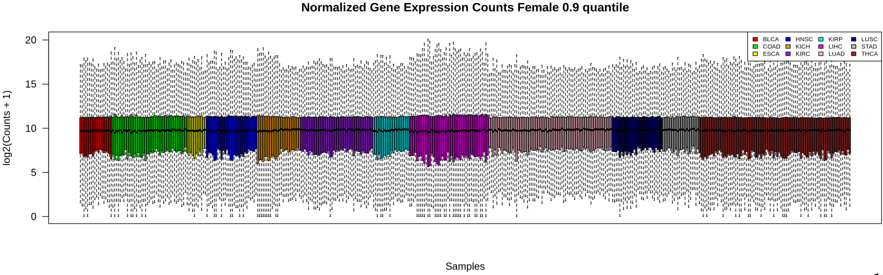
Sample A
5 genes expressed at same level
20 total reads; 4 reads/gene

Sample B
5 genes expressed at same level as sample A
2 highly expressed genes represent 50% of the 20 reads

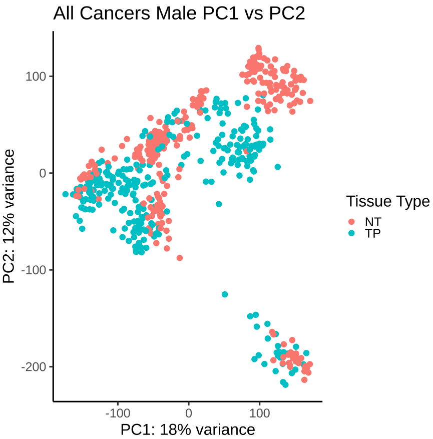
A few highly differentially expressed genes between samples, differences in the number of genes expressed can skew some types of normalization methods

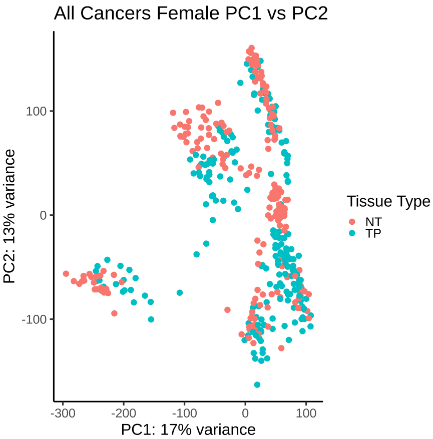
Normalizing Pan-Cancer Expression Matrices





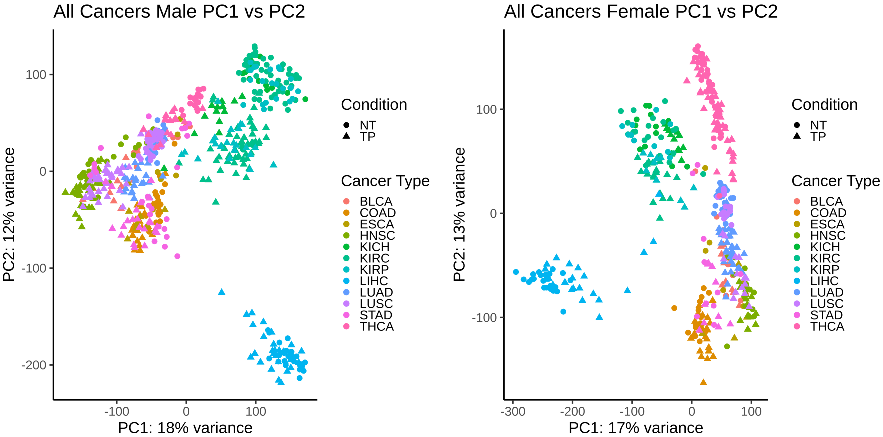
Normal and Tumor Tissue Differ ACROSS 12 Cancer Types





Across all 12 cancer types, the observed variation between healthy and tumor tissue samples is unlikely to have occurred by chance alone (NPManova p-val = 0.0001)

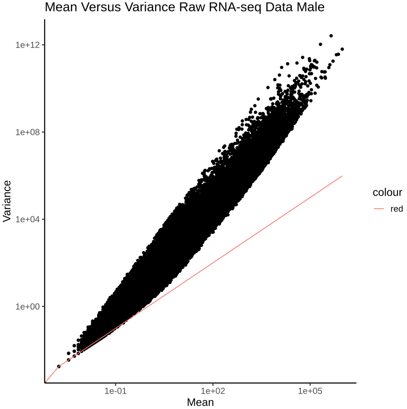
Normal and Tumor Tissue Differ WITHIN each Cancer



Within each cancer type, the observed variation between healthy and tumor tissue samples is unlikely to have occurred by chance alone (Acceptions: Males: ESCA (LOW SAMPLE SIZE), Females: BLCA, ESCA (LOW SAMPLE SIZE), STAD (LOW SAMPLE SIZE))

Why does DESeq2 Use the Negative Binomial Distribution

- Reads are count based hence they cannot be normally distributed
- Variance tends to be greater than the mean, especially for genes with large mean expression values
- We need to account for this increase in variance using the Negative Binomial model



DESeq2 Accounts for Increased Variance

1. Estimate gene-wise dispersions: captures biological variability in gene expression across samples
2. Fit negative binomial GLM to count data: This model incorporates the estimated gene-wise dispersions as a parameter to account for variability in the data
3. Use fitted GLM to identify differentially expressed genes

The use of linear models allows for more complex designs

$$design = \ cancertype + condition$$

This tells DESeq2 to test the effect of condition while controlling for the effect of cancer type

Finding Differentially Expressed Genes (DESeq2)