**PROJECT PART I: Inter-genomic expression analysis**

**Purpose:**

Assesses whether the selection of this particular set of genes could make a god set to base the analysis on

**Questions**:

A) Cancerous

1. **Does the aggregate expression of the genes in the set of interest in the data differ from the expression of other genes not in the list (i.e control set)?**

* What is the typical distribution of the data (difference between the general expression levels)?
* Are there any significant differences in the data when selecting between the different control groups ?
* How similar is the data from the different control groups?
* For each control set - SoI pairing is the difference significant?

1. **Does the expression difference remain true for all Tissue types** **?**

* What does the difference in expression look like when the data is parsed by Tumour source type?
* Does the difference remain significant?

B) Non-cancerous

1. **Does the aggregate expression of the genes in the set of interest in the data differ from the expression of other genes not in the list (i.e control set)?**

* What is the typical distribution of the data (difference between the general expression levels)?
* Are there any significant differences in the data when selecting between the different control groups ?
* How similar is the data from the different control groups?
* For each control set - SoI pairing is the difference significant?

1. **Does the expression difference remain true for all Tissue types** **?**

* What does the difference in expression look like when the data is parsed by Tissue type?
* Does the difference remain significant?

**Data**:

* TCGA cancerous RNA TPM data: 10535 samples x 60499 genetic identifiers (genes), samples are collected from a variety of projects
* GTEx non-cancerous RNA TPM data: 15201 samples x 56200 genetic identifiers (genes), samples are taken from individuals post-mortem

**Components**:

The TCGA cancer data was cleaned such that: non-cancerous samples were removed, samples that were not human based (cell lines) were removed

The GTEx non-cancer data was cleaned such that: diseased samples were removed, samples which did not meet a certain standard of RNA integrity (RIN) were removed.

For each of the TCGA dataset, the following was performed:

* The aggregate measurement (mean) of the log2 expression levels of the genes in the set of interest was taken.
* Creation of multiple control sets made up from a random selection of genes excluding those that appear in the set of interest (leftover genes). The number of genes equals that of the genes in the set of interest.
* The number of control sets is variable, terminating only once every gene in the list of leftover genes has been incorporated into a control set at least once. A control set cannot only contain genes that have already been included in other sets. Therefore the upper limit is set at the number of genes in the leftover list (. For the current run the number of controls attained 547.
* For each control set, the aggregate measurement (mean) of the log2 expression levels of the control genes was taken.
* For each control set, the difference between the the set of interest’s expression mean and it’s own was calculated

Analyses:

**Q1. A**: Violin Plot Distribution

* **Type**: Graph
* **Description**: Violin plot of the **distribution of the mean expression difference between the mean expression within the SoI vs Control set for multiple Control sets**
* **Purpose**: Provides **insights into the distribution type (non-normal) which is necessary to determine the types of statistical tests** that need to be performed and provides insights into the pattern/variation of the distribution using different control sets
* **Interpretation**: TCGA plot: Indicate the non-normal distribution of several of the mean differences, indicating that the use of **non-parametric tests may be best.** Additionally, each of the violin plots shows long negative tail (negative skewness) which adds to the non-normality of the distribution but also indicates that some of the differences between the set of interest and the control gene-set were closer together. This might also be expected given that we are dealing with cancer samples in which there is a greater variation and can therefore be over-expression of all genes. GTEx plot: Indicate a more normal distribution compared to those of the TCGA data.Each of the violin plots show long negative tail (negative skewness) but also a positive tail (not as long), indicating that the differences between the set of interest and the control set can be closer together or far apart more typical of what would be expected or a truly random phenomenon and given that these samples are non-cancerous.

**Q1.B**: Friedman test

* **Type**: Test
* **Description**: Friedman’s test (non-parametric version of the Repeated Measures ANOVA, RMA) on t**he mean expression difference between the SoI vs Control set for all Control/SoI pairings**
* **Purpose**: assess whether there is a **significant change in the SoI/Control expression pattern when using different pairings** (i.e. testing the similarity of all control groups)
* **Result**: The Friedman Test turned out significant for both the control sets originating from the TCGA data as well as those from the GTEx data both giving a value of 2.2e-16 indicating that at least one of the Set-of-interest to control-set pairings demonstrated a significantly different distribution.
* **Interpretation**: The Friedman should be expected to turn out significant as the different control sets are all unique in their genomic/transcriptomic composition and more so for the cancerous data which would show more variation in the expression of genes.

**Q1.C**: Contrast Analysis **NOT IMPORTANT**

* **Type**: Test
* **Description**: Two-samples Wilcoxon test (non-parametric version of the paired t-test) on t**he mean expression difference between the SoI-Control\_set pairing and the global mean expression difference of all other pairings for each pairing**
* **Purpose**: assess whether there is a **significant change in any of the pairings compared to the global pattern** this allows for the rejection that the general expression of the control groups are all very much distinct
* **Result**: The Contrast analysis shows that the expression pattern of the control sets seem to differ significantly compared to their average with only 7/547 showing a p-value > 0.05 for the paired Wilcoxon Signed Rank Test. Cannot therefore assume that all the control sets behave in the same general fashion. This result was both achieved for the cancerous TCGA dataset and the GTEx non-cancerous dataset.

**Q1.D**: Pairwise Boxplot of expression difference

* **Type**: Graph
* **Description**: Boxplot of the **difference between the mean expression within the SoI vs Control set for all Control sets**
* **Purpose**: Shows the trend of the difference between the SOI and Control-Set thereby assessing if the distributions share the same general characteristics despite the Friedman test showing up as significant
* **Interpretation:** Despite that fact that the Friedman Test and the Contrast Analysis resulted in significant results (indicating a significant difference in differential expression between the pairings), the plot hints at the difference in expression between every control set and the Set-of-Interest being bound in a consistent range centred between 8 and 10 on a log2 scale. And that the difference between the general expression of both sets is consistently positive, demonstrating increased expression within the SoI compared to all control sets. This pattern is maintained when using the GTEx non-cancerous dataset as well as when using the TCGA dataset

**Q1.E**: Paired Permutation Test

* **Type**: Test
* **Description**: **Two-sided** **Paired permutation test between the SOI and the control set for each control group corrected using the Benjamini-Hochberg correction** method for multiple testing (due to large number of tests)
* **Purpose**: Gives evidence as to whether there is a significant difference between the expression levels of the SoI vs controls for all control sets
* **Result**: The results show a p-value of 0 for all pairings. The result is the same when using the TCGA data or the GTEx data proving that this difference in expression is not an artefact of cancer.

**Q1.F**: Boxplot of the expression for the SoI vs 3 control groups

* **Type**: Graph
* **Description**: Boxplot of the **distributions of the expression levels for the SoI and 3 control sets**
* **Purpose**: Demonstrates the difference in the general expression pattern in clear fashion
* **Interpretation**: the p-value and plots demonstrate that the expression pattern of the control sets are significantly different from that of the SoI. The result is the same when using the TCGA data or the GTEx data proving that this difference in expression is not an artefact of cancer.

**IMPORTANT STEP**: Can now decrease of the number of control sets which to a single control set for future use

**Q2. A**: Tumour specific (TCGA) expression difference Barplot (Tissue specific for the GTEx data)

* **Type**: Graph
* **Description**: Facetted bar plot of the **difference in mean expression between the SoI vs Control set for each sample by Tumour site (for TCGA) and then by Tissue (for GTEx)**
* **Purpose**: Visually **demonstrates whether the SoI-to-Control-set expression pattern is maintained across all tumour or tissue types** and allows assessment of any obvious cancer-specific or tissue-specific variations
* **Interpretation**: Indicates that the expression pattern of positive ∆ mean expression between the SoI and the Control set is maintained across all cancer types, without much variation in the pattern. This is both true for the TCGA data as well as the GTEx data.

**Q2. B**: Tumour-specific / Tissue-specific permutation test

* **Type**: Test
* **Description**: **Two-sided** **Paired permutation test between the SOI and the control set for each Tumour site corrected using the Benjamini-Hochberg correction** method for multiple testing (due to large number of tests)
* **Purpose**: Gives evidence as to whether there is a significant difference between the expression levels of the SoI vs controls for for each cancer type
* **Results**: All of the tumour types are shown to have a significant difference between the general expression of the SoI and the Control set

**Q2. C**: Average expression difference by tumour type (for TCGA) or Tissue type (for GTEx) Barplot

* **Type**: Graph
* **Description**: Bar plot of the **mean** **difference in general expression between the SoI vs Control set by Tumour type (for TCGA) or Tissue Type (for GTEx)**
* **Purpose**: Visually **demonstrates the general expression pattern based on the different tissue types and tumour types**
* **Interpretation**: Indicates that the expression pattern of positive ∆ mean expression between the SoI and the Control set is present across all cancer types and that there are some differences in these levels

**Results**:

Cancerous (TCGA):

* The genes in the set of interest differ significantly in their global mean expression compared to any random selection (set) of genes from the same cancerous samples .
* The over-expression is maintained across all SOI/Control-Set pairings and the global expression difference remains in a narrow range of 8-10 with an average of around 9 on a log2 scale, signifying an increase of about 512 times. Despite this the Friedman Test (non-parametric RMA) as well as the contrast analysis showed significant p-values indicating that each control set is also unique in its expression profile.
* All Tumour types display a positive mean expression difference (set of interest - control set) that is significant between the set of interest and the control set.

Non-Cancerous (GTEx):

* The genes in the set of interest differ significantly in their global mean expression compared to any random selection (set) of genes from the same non-cancerous samples .
* The over-expression is maintained across all SOI/Control-Set pairings and the global expression difference remains in a range of 7-10 with an average of around 8 on a log2 scale, signifying an increase of about 256 times. Despite this the Friedman Test (non-parametric RMA) as well as the contrast analysis showed significant p-values indicating that each control set is also unique in its expression profile.
* All Tissue types display a positive mean expression difference (set of interest - control set) that is significant between the set of interest and the control set.

**Interpretation**:

* The heterochromatic/LAD genes of interest display a greater average over-expression compared to the average most other genes throughout the genome in cancer samples that is reflected in non-cancerous samples.
* The uniformity in behaviour (all being positive and in a similar range) reinforces the robustness of the expression pattern of the gene set of interest. It suggests that this upregulation is a stable characteristic across various control comparisons and neither a random artefact due to a specific selection of control genes or an oncogenic phenomenon.
* Yet the analysis still reveals this trend which would be intriguing to analyse in more detail in terms of the individual genetic expression profiles.