**Inferring persistent molecular changes in airways**

**associated with lung cancer in former smokers**

Liam Brockley

MSc student, Integrative Oncology

[lbrockley@bccrc.ca](mailto:lbrockley@bccrc.ca)

Dr. Wan Lam

[wanlam@bccrc.ca](mailto:wanlam@bccrc.ca)

**Background:** Former tobacco smokers remain at high risk for lung cancer long after they quit smoking. Previous studies have found that normal-appearing airway cells from former smokers retain some of the altered gene expression patterns that are measured in current smokers. If these persistent changes are shown to contribute to lung cancer, and we can find the mechanisms that cause them to persist, they could be targets for early detection, screening, or interception. In my project, I compare multiple datasets to infer which genes are persistently altered, whether DNA methylation (a type of epigenetic mark) can explain why they remain altered, and whether they could be linked to lung cancer.

**Hypothesis**: We can infer which persistent smoking-induced gene expression changes in former smokers’ airways could contribute to carcinogenesis of lung adenocarcinoma.

**Rationale**: These persistent gene expression changes could be intercepted before lesions form, used to stratify former smokers by cancer risk for screening programs, or detected in a non-invasive manner.

**Experimental Design:** In the first main part of the analysis applied to lung adenocarcinoma (LUAD), I am using the following 4 datasets:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset Identifier** | **Description** | **Number of samples** | **Cohort ID** | **Platform** |
| A1 | Expression data of small airway brushing samples from healthy current and never smokers | 182 (112 CS,  70 NS) | GSE63127 | GPL570  microarray |
| TE | Expression data from paired LUAD tumor and adjacent normal samples | 114 (57 T-N pairs) | TCGA-LUAD  expression data | Illumina  RNA seq |
| A2 | Expression data of bronchial brushing samples from healthy current, **former**, and never smokers | 104 (52 CS, 31 FS, 21 NS) | GSE7895 | GPL96  microarray |
| TM | Methylation data from paired LUAD tumor and adjacent normal samples | 58  (29 T-N pairs) | TCGA-LUAD methylation data (cbioportal) | Illumina  HM450K |

I wish to compare these datasets to each other to determine which genes could be altered in airways due to smoking, persistently altered after quitting, altered in lung adenocarcinoma, and altered due to DNA methylation. For genes found to be relevant in all four datasets, we could make the following tentative inference:

**A close-up of a sign

AI-generated content may be incorrect.**

**Statistical Questions:**

For the first section of the first part of my project, I am trying to define a list of genes that may be relevant both to the short-term effects of smoking in airways and to lung cancer. I want to find a way to determine whether the list of genes I define is likely to be relevant to both processes or whether it is likely to be discovered due to chance. For my most recent approach, I am using the first two datasets shown above, and filtering them with the steps shown below:

A diagram of a number

AI-generated content may be incorrect.

To review the terminology used here:

* A1 refers to genes that are differentially expressed between current and never smokers in small airway brushing samples.
* TE refers to genes that are differentially expressed between matched tumor and adjacent normal samples from lung adenocarcinoma (LUAD).
* The numbers in the boxes represent the number of genes remaining after each filtering step.
* FDR refers to false discovery rate, which is an adjusted P-value accounting for multiple comparisons.
* log2FC refers to log2 fold change, a measure of the difference between two groups:

log2​FC=log2​( )

For A1, log2FC measures the difference between average normalized expression levels for current vs. never smokers. For TE, log2FC measures the difference between average normalized expression levels for paired tumor vs. adjacent normal samples. When I filter for the same sign in step 6, I filter for genes that are altered in the same direction in current smoker airways compared with never smoker airways, as they are in tumor tissue compared with adjacent normal tissue.

I am trying to use the hypergeometric test to check whether the number of genes that are shared between datasets in these filtering steps is significantly different from what would be expected by chance. Strangely, my test seems to show that the 3.7 k genes overlapping after filter step 3 are no more than expected by chance, yet the 450 genes overlapping after step 5 are much more than expected by chance. The R code is shown below:

*Testing filter step 3:*

N = 14 941 # All shared gene symbols

m = 5899 # A1 genes with FDR < 0.05

n = 9508 # TE genes with FDR < 0.05

k = 3755 # Observed overlap

p\_value <- phyper(k - 1, m, N - m, n, lower.tail = FALSE)

**p\_value**: 0.49

**Interpretation**: The observed overlap is not significantly different than expected by chance.

*Testing filter step 5:*

N = 3755 # All genes with FDR < 0.05 in both airway and tumor

m = 940 # Airway genes with FDR < 0.05 in both airway and tumor, and top 25% |log2FC|

n = 940 # Tumor genes with FDR < 0.05 in both airway and tumor, and top 25% |log2FC|

k = 450 # Observed overlap

p\_value <- phyper(k - 1, m, N - m, n, lower.tail = FALSE)

**p\_value**: 5.77e-72

**Interpretation**: The observed overlap is significantly greater than expected by chance.

This result seems counterintuitive. My interpretation is that genes that have FDR < 0.05 in A1 and TE individually only have FDR < 0.05 in both datasets due to chance, yet if we take the top n % of genes from each dataset as defined by log2FC, the resulting overlap of genes is significant. In fact, this overlap at step 5 is still significant even if instead of taking the top 25% of genes, I take the top 96% of genes (i.e. 4th percentile: p-value = 9.2e-4). I am unsure if I am choosing, performing, or interpreting this methodology correctly, and I would like to seek advice regarding this matter. To reiterate my goal, I wish to define a list of genes that could play a role in both the short-term effects of smoking in airways and in lung adenocarcinoma, and to test how likely these genes are to truly be relevant to both processes to a degree higher than chance.