## Slide 1: Title

**Differential Methylation in Primary vs Metastatic Cancer**

The Splice Girls

Almas Khan, Denitsa Vasileva, Diana Lin, Nairuz El-azzabi

## Slide 2: Motivation - 1 min

Late-stage Head and Neck Squamous Cell Carcinoma (HNSCC) is associated with high mortality rates. To improve this, it is crucial to develop methods to differentiate between HNSCC lung metastases and primary squamous cell carcinoma (LUSC)- a different, highly-treatable type of cancer that is frequent in patients with HNSCC

Previous studies have identified differentially methylated gene loci implicated in primary vs metastatic tumours in other cancer types

The aim of our project is to identify differentially methylated CpGs which can be used to distinguish between HNSC lung metastases and primary LUSC. This will allow us to gain greater understanding of the differences in the genetic and epigenetic mechanisms between the two types of cancer which can aid in the development of tools for more accurate differentiation and diagnosis.

## Slide 3: Question - 30 sec

Our research question was are there any CpG sites differentially methylated in primary compared to metastatic cancer. This is an important question, as the differentiation between the two types of cancer has a big impact in treatment and prognosis but is currently difficult to accomplish.

Given the limited amount of primary HNSC samples (3) and the results we got from our PCA analysis we decided to combine it with the lung primary cancers, to compare primary cancers to the metastatic HNSC and see if there was any differential methylation between the 2.

## **Slide 4:** **Dataset - 1** min

The dataset comprises 51 male samples of various ages (45-80 years) with either HNSC or LUSC. Tumour tissue Methylation was measured using the Illumina EPIC array which measures methylation at ~850K CpG sites using fluorescence labelling coupled with bisulfite conversion. The data was already normalized using noob normalization. Noob-normalization is used on array values to account for bias in the fluorescence and dye. It uses a normal-exponential convolution model to estimate the true signal conditioned on observed intensities. These values are then normalized for intensity differences between the dyes using a multiplicative scale factor.

The Beta value represents the ratio of intensities of methylated over total intensity (sum of methylated and unmethylated values) and is usually between 0 and 1. M values are the log transformed ratio of methylated over unmethylated probe.

## Slide 5: Workflow - 5 sec

This is the overview for our workflow. We couldn’t get gene set enrichment to work with ermineR so we left out that step.

## Slide 6: PCA Table - 1 min

To see if there is an underlying signal in our data, principal component analysis was conducted, using the beta values. This table is the summary table from the pcs object, i.e. summary(pcs). The first row, standard deviation, contains the eigenvalues, as our values have already been centered and scaled using t(scale(t(data)). The eigenvalues mean that our components have been sorted from largest to smallest based on their standard deviation. The second row, proportion of variance, shows the amount of variance that each principal component accounts for. In our table, it is evident that PC1 accounts ~15% of the total variance in the data, which is on the lower side for the first principal component. The third row, cumulative proportion, is the cumulative amount of explained variance. Using this value, you can see that using the first 10 components only accounts for ~55% of the total data.

## Slide 7: PCA Clustering Plots - 45 sec

As per standard procedure, we plotted PC1, PC2, PC3, which cumulatively account for ~32% of the variance in our case. PCA did not show clustering of cancer type (primary vs metastatic) or cancer origin (HNSC or LUSC). In attempts to remedy this, the data was rescaled and recentered, to no avail. Similar plots using M values (instead of beta values) were also generated, yielding similar results. Since the data was already normalized from the GEO dataset, we attribute this ‘non-clustering’ to the fact that we didn’t do the normalization ourselves (it was pre-done). If we could do it all over again, we would run minfi using the raw .idat files to see if this issue persists.

## Slide 8: Box Plot Distribution (by cancer\_type) - 20 sec

Given that there was no clustering observed in the PCA analysis, a box plot was generated to visualize the distribution of beta values across the different samples, to see if any patterns are detected. Looking at the boxplot, the mean varies within each group, as well as between the two groups, with no clear pattern discerned.

## Slide 9: Box Plot Distribution (by cancer) - 15 sec

The same box plot was re-generated, this time colouring by the 4 different cancers, but no pattern surfaced in this case either. This is again attributed to the fact that the normalization was pre-done, and perhaps was done improperly, or the chosen normalization method (noob) did not suit our needs.

## Slide 10: Beta Value Density Plots - 30 sec

The Beta Value Density plot is an exploratory visualization tool. The density plot can be used as a Quality Control check to show the distribution of beta values, to compare their distribution across different conditions (e.g. two types of cancer) and to point to a need for normalization of the values. Beta distribution is along the [0,1] interval.

In our case, plots show consistent beta values distribution between the two types of cancers. This was the expected result as the data had been normalized. In addition, there are two peaks- close to each end of the distribution.

## Slide 11: P-value histogram - 1 min

Limma, eBayes and topTable were used to generate a test for differential methylation for each CpG probe/site in which the tested null hypothesis is very similar to that of the *t*-test.This p-value histogram serves as a good sanity check to perform hypothesis testing of whether methylation signal is affected by experimental variables/conditions such as cancer\_type and age.

1- when performing hypothesis testing of whether CpG methylation is affected by the interaction between cancer\_type (primary vs metastatic) and age (right graph), p-values are somehow distributed uniformly with very high “depth” meaning that there are alot of p-values that are null and probably a small percentage of alternative p-values. An FDR correction method such as Benjamini-Hochberg can help with identifying those significant alternative hypotheses. In total, we can say that there is no sufficient evidence to reject the null hypothesis.

2- when performing hypothesis testing of whether CpG methylation is affected by cancer type only (left graph), there seems to be more small p-values that appear as a peak near 0 representing the alternative hypothesis implying that the analysis has some statistical power to reject the null hypothesis.

## Slide 12: P-value density plot - 45 sec

DNAm-widehypothesis testing was performed using a 2-sample t-test and a density plot of p-value distribution was generated using geom\_density()*.* T-test performs hypothesis testing as a part of a quality control check to assess whether CpG methylation in this case is affected by cancer type(primary vs metastatic). According to the graph, there seems to be more small p-values that are located near 0 where the alternative hypotheses are located. This is very similar to the right graph above showing that there is some statistical power to reject the null hypothesis and conclude that cancer type (independent variable) can correlate with the changes seen in CpG methylation signal (dependent variable).

**Slide 13: Hierarchical clustering heatmap - 15**

The hierarchical clustering was done using a Euclidean distance and average linkage methods. The clustering heatmap variables are shown above. As you can see, there was no obvious pattern of clustering within the heatmap, similarly to the PCAs conducted.

**Slide 14: Statistical Analysis - 25 sec**

Limma package was used to test for probe-wise differential methylation as it performs linear regression on M-values to assess CpG differential methylation in the context of multifactorial design experiment (~cancer\_type\*age). Limma accounts for multiple hypothesis testing, and biased variance estimates using a moderated t-test by using the function eBayes()and topTable() to make multiple testing adjustments to the statistical significance measures. Other statistical methods were also used as mentioned earlier including p-values and t-statistics to screen all possible hypotheses and find the ones that are statistically significant.

**Slide 15: Limma details - 30 sec**

The statistical analysis was done using an interaction model with cancer\_type and age. Age was included as a covariate due to changes in methylation that can solely occur over a normal lifespan and because we had wide ranges of ages. We account for sex differences by only using the larger group of male samples, as there was only a small group female samples so we did not use them. The default Benjamini-Hochberg multiple test correction was used.

## Slide 16: TopTable-coefficient of primary cancer - 20 sec

Here were the top genes in the limma analysis using the coefficient of primary cancer. As you can see, couldn’t find any significant adjusted p values for any of these genes.

We used the annotation provided from Zhou et al (2017).

## Slide 17: Strip Plot of Top 10 Differentially Methylated Genes - 20 sec

The strip plot depicted shows the beta values of the probes corresponding to the top 10 differentially methylated genes, coloured by cancer\_type. In this plot, you can see there are some genes that are positively differentially methylated in primary cancer whereas other genes are positively differentially methylated in metastatic cancer. The black line represents the mean of each group. One probe (NA) had no annotation.

## Slide 18: Chromosome Plot - 20 sec

From each differentially methylated probe, the chromosome coordinate was extracted from the annotation and plotted along the respective chromosome length. The plot depicts the differentially methylated sites along each chromosome. It is a visual representation of the differentially methylated regions of each chromosome in primary cancer.

## Slide 19: Pathway Analysis KEGG - 30 sec

The top genes from the topTable analysis were used to perform pathway enrichment analysis using the KEGG database. KEGG was chosen as it provides detailed, curated pathway annotations. In total, there were 45 implicated pathways. The table illustrates the top eight pathways associated with our genes of interest. It is of note that many of these pathways have previously been implicated in cancer. For example, alcohol consumption has been implicated in the literature as a risk factor for head and neck cancer. In addition, one of the implicated pathways is viral carcinogenesis. Indeed, HPV- a viral carcinogen- has been implicated in HNSCC.

Gene Ontology terms were also obtained for each top gene, but we were unable to perform gene set enrichment analysis using the package ermineR due to software incompatibility.

**~9.8 min**

## Slide 20: Discussion - 20s

We didn’t find significant adjusted p values for primary vs metastatic. We also couldn’t find any differential methylation when looking at the interaction term or age. However, given our small sample size, statistical power may be limited. It would be interesting to see if we get similar results using a much larger sample size for each group, if we had the computational resources.

## Slide 21: Challenges - 1 min

One of the main challenges that we faced during our analysis was the large data size. Our dataset consisted of 52 samples of ~850K CpG sites each. This presented problems with lack of computational resources and computation speed which made analyzing the data challenging. In the future, this problem can be resolved through the use of a cluster, or selection of a smaller dataset when faced with a lack of computational resources.

Another challenge that we faced was the unbalanced sample size of the different cancer subtypes within the data. For example, in the data there were 3 primary HNSC samples and 3 Primary lung squamous cell carcinoma samples compared to 16 Primary Second Squamous Cell Lung Carcinoma and 28 Metastatic HNSC samples.

This was a problem as it prevented us from doing more in-depth analysis by cancer subtype because there were not enough samples available for each and the results would not have been meaningful. This problem could have been resolved by a more balanced experimental design of the initial study or by the combination of multiple datasets from multiple studies/

Finally, technical challenges included incompatibility between our java installations and the ErmineR package as well as lack of clear, consistent labeling standard for the metadata in our dataset.

## Slide 22: Conclusions - 15 sec

There was not enough statistical power to conclude that there is a differential methylation of CpG sites/islands between primary (HNSC and LSCC) and metastatic (pulmonary HNSC) cancer. This is based on the results obtained from limma and topTable() as the top differentially methylated genes did not have significant adjusted p-values. This could be attributed to an unbalanced dataset and small sample size.