Differential Methylation in Primary vs Metastatic Cancer

The Splice Girls

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

Motivation

- **Late-stage Head and Neck Squamous Cell Carcinoma (HNSCC)**:
  - It is associated with high mortality rates
  - It can metastasize to the lung and causes **HNSCC lung metastases**
- **Lung Squamous Cell carcinoma (LUSC)**:
  - It is a type of a primary lung cancer
  - It is high-treatable
  - It is frequent in patients with HNSCC

Aim: Identify differentially methylated CpGs between **HNSCC lung metastasis** and **primary LUSC** → gain better understanding of epigenetic and genetic differences between the two types → more accurate diagnosis and differentiation

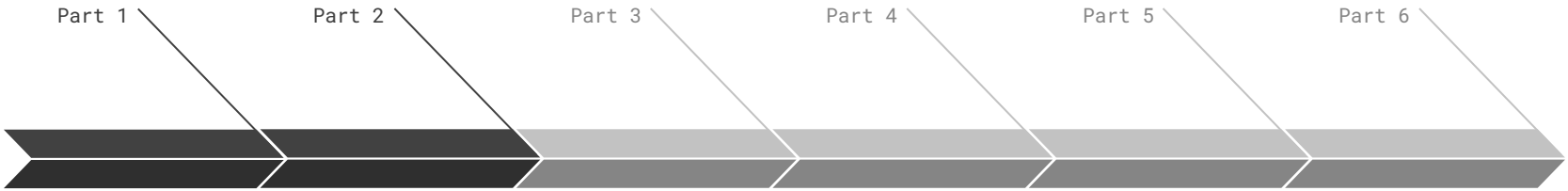
Research Question

Are there individual CpG sites/islands with significant differences in methylation level in primary vs metastatic cancer?

About the Dataset

GEO Accession	GSE124052
Method	Infinium MethylationEPIC BeadChip
Values	Raw and normalized Beta values
Participant Profile	51 males (45 to 80 years old) with either LUSC or HNSCC

Overview of Workflow



Gather the Data

Wrangle the data

Reshape the data

Filter the data

Combine with metadata

Quality Control

PCA

P-value Histogram

Data Exploration

Beta value boxplots

Beta value density plots

Hierarchical clustering

Heatmap

Statistical Analysis

Convert beta values to M values

Limma models

Annotation of CpG sites

Visualizing Results

Beta value strip plot of differentially methylated genes

Chromosome plot of differentially methylated genes

Pathway Analysis

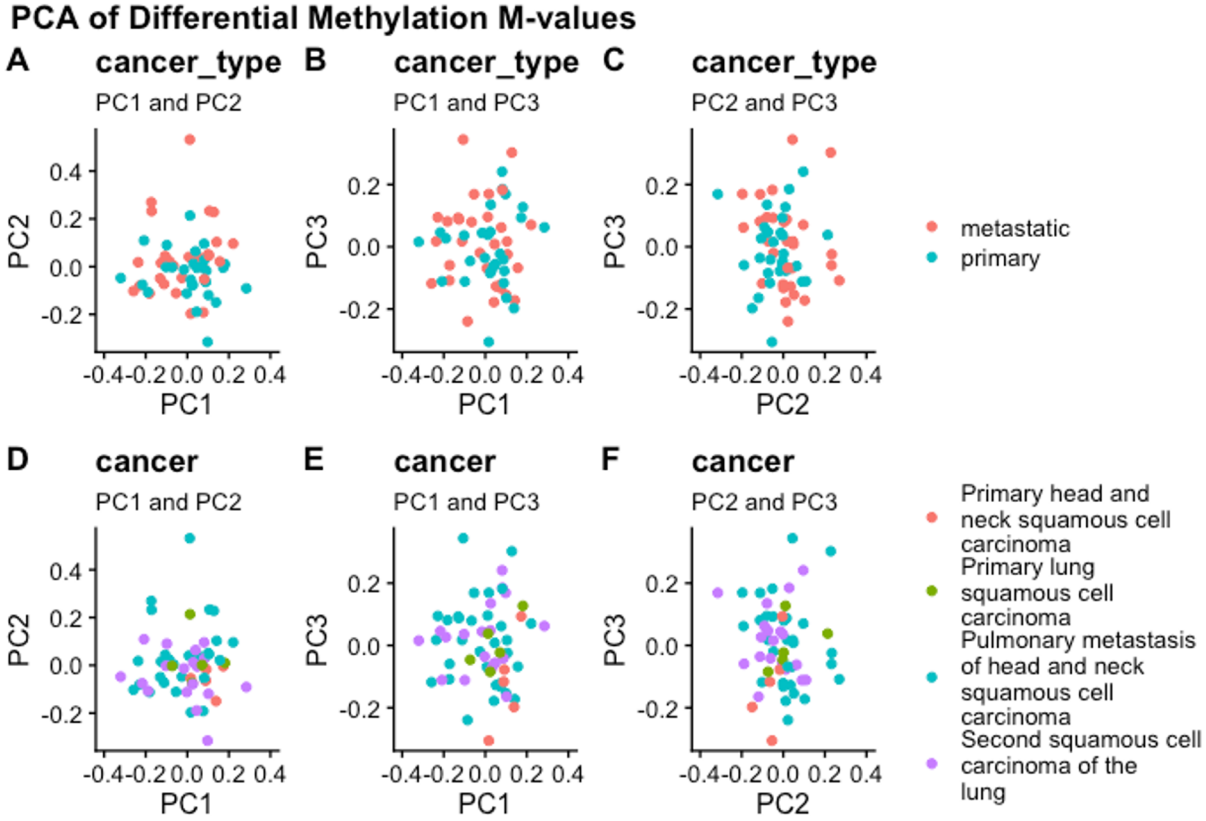
Annotation of differentially methylated genes using gene ontologies

KEGG Pathway annotation of differentially methylated genes

Quality Control: Principal Component Analysis

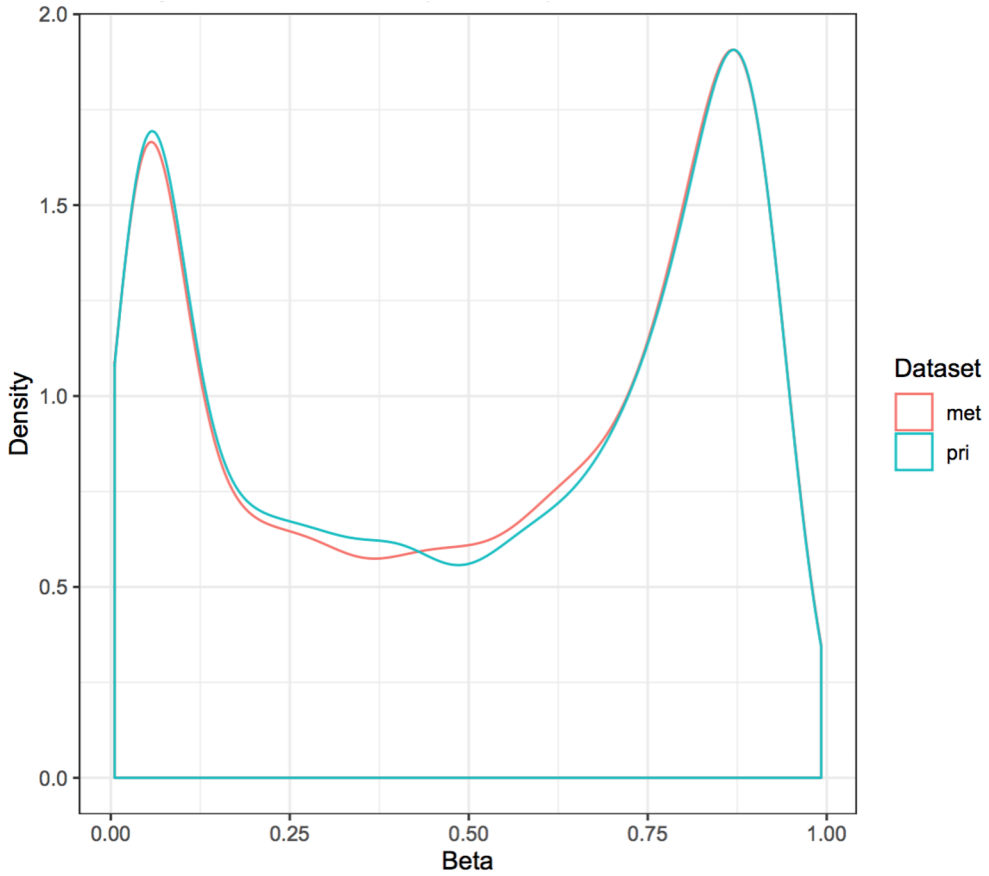
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	3.0702	2.4583	1.94006	1.82127	1.56238	1.44677	1.36316	1.3169	1.24193	1.17200
Proportion of Variance	0.1571	0.1007	0.06273	0.05528	0.04068	0.03489	0.03097	0.0289	0.02571	0.02289
Cumulative Proportion	0.1571	0.2578	0.32056	0.37584	0.41652	0.45141	0.48238	0.5113	0.53699	0.55988

Quality Control: Principal Component Analysis



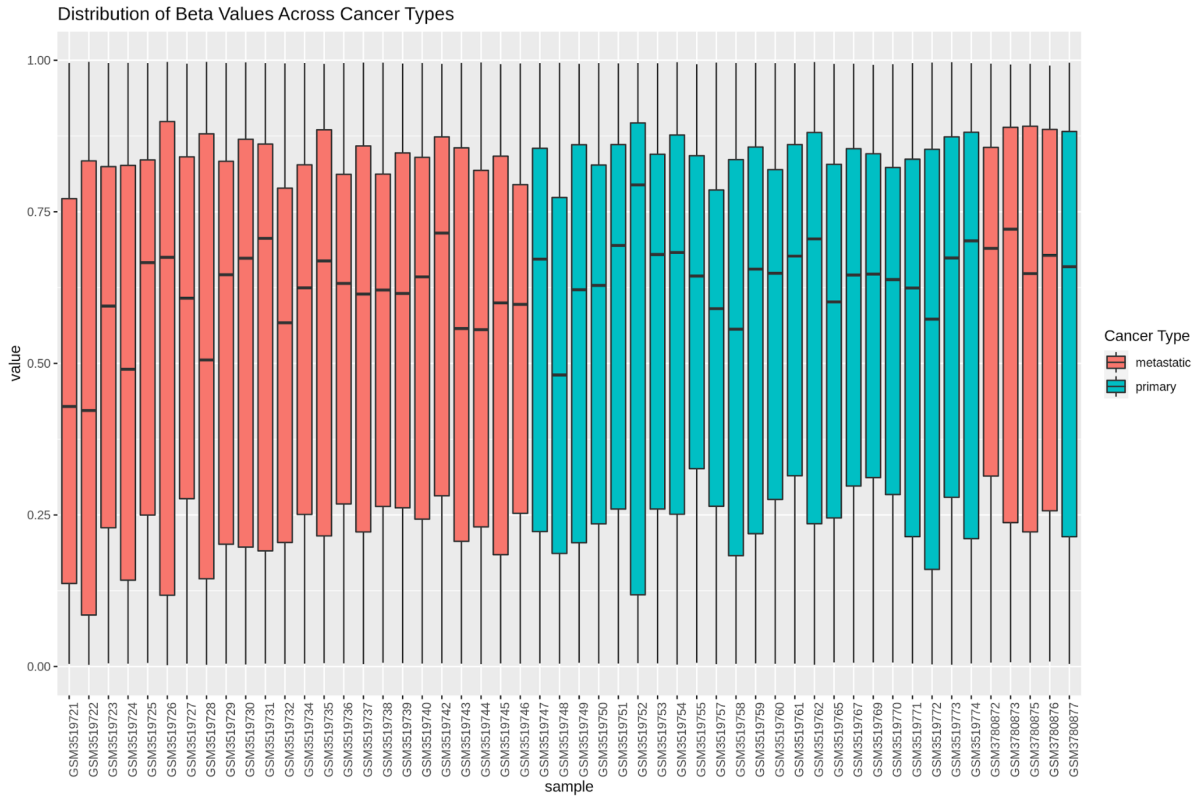
Quality Control: Density Plot of Beta Values

Density Plot of Beta Values Across Primary And Metastatic Cancer



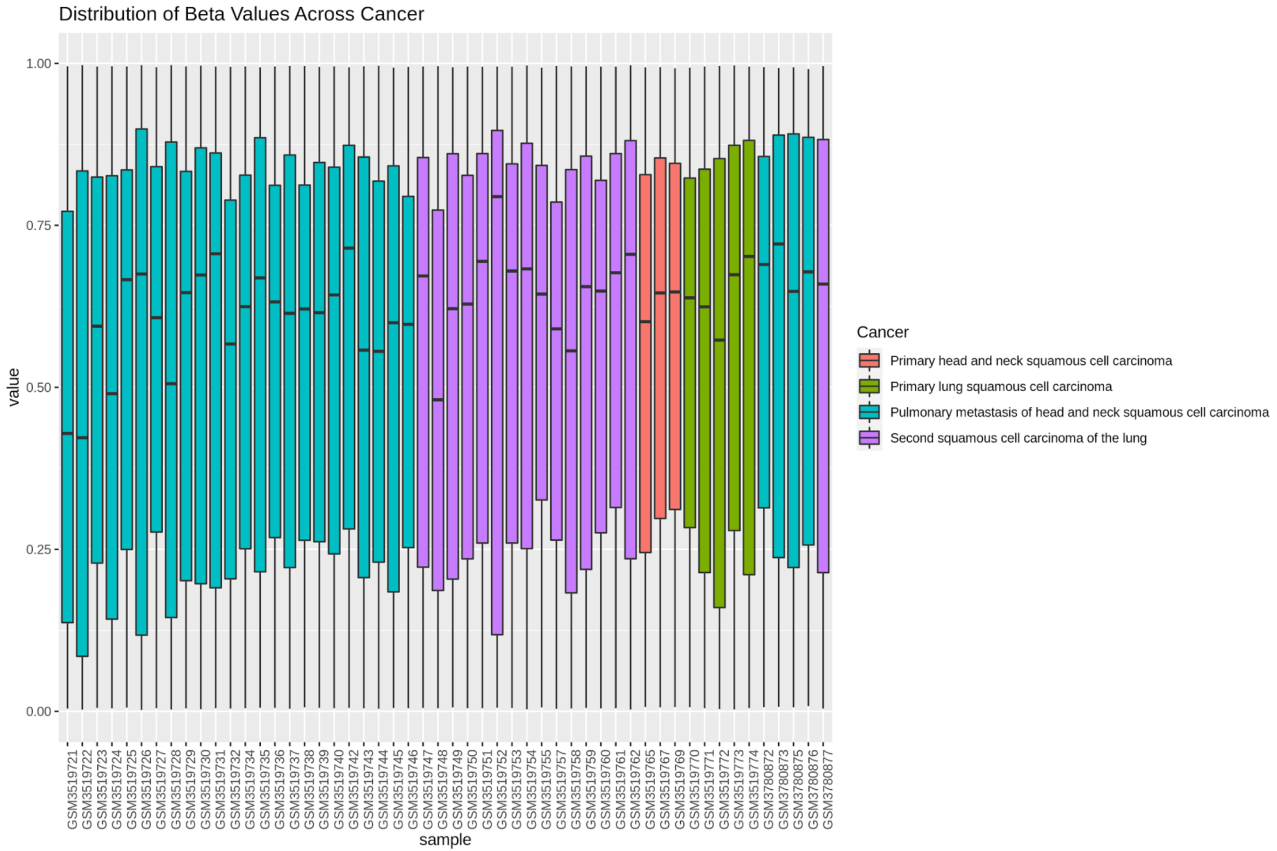
Quality Control: Box Plot Distribution

Beta Value Boxplot Distribution Across Cancer Types



Quality Control: Box Plot Distribution

Beta Value Boxplot Distribution Across The Cancers



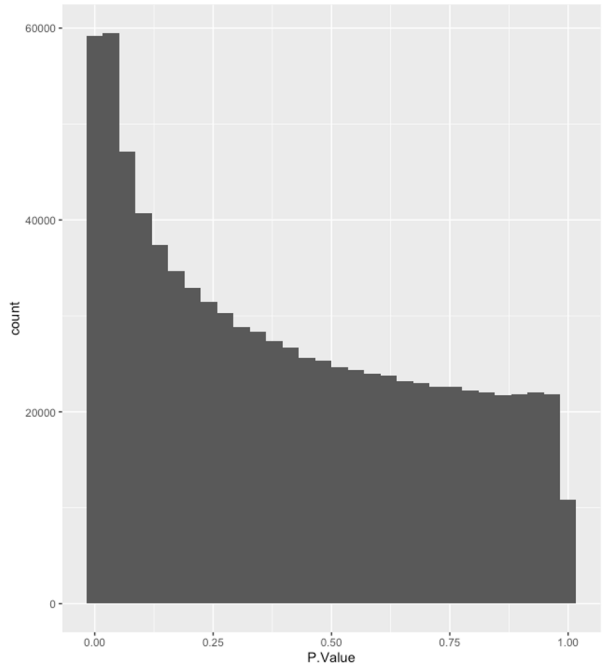
Quality Control: p-value histogram

```
#design the model
design4 <- model.matrix(~cancer_type , limma4_metaData)

#fit the model
fit4 <- lmFit(limma4_expData, design4) %>% eBayes()

#---------------------------------p-value histogram distribution
pvalues <- topTable(fit4, number = Inf)
```

```
pvalHistogram4 <- pvalues %>%
  ggplot (aes(P.Value)) + geom_histogram()
```

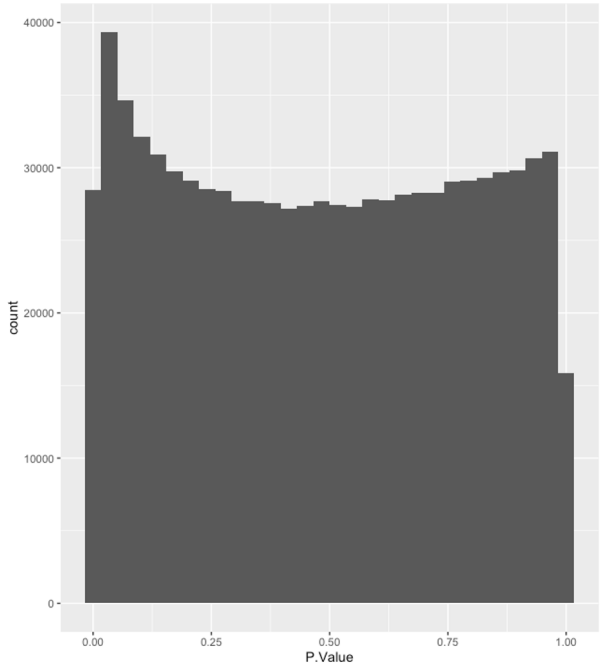



```
#design the model
design4 <- model.matrix(~cancer_type*age , limma4_metaData)

#fit the model
fit4 <- lmFit(limma4_expData, design4) %>% eBayes()

#---------------------------------p-value histogram distribution
pvalues <- topTable(fit4, number = Inf)
```

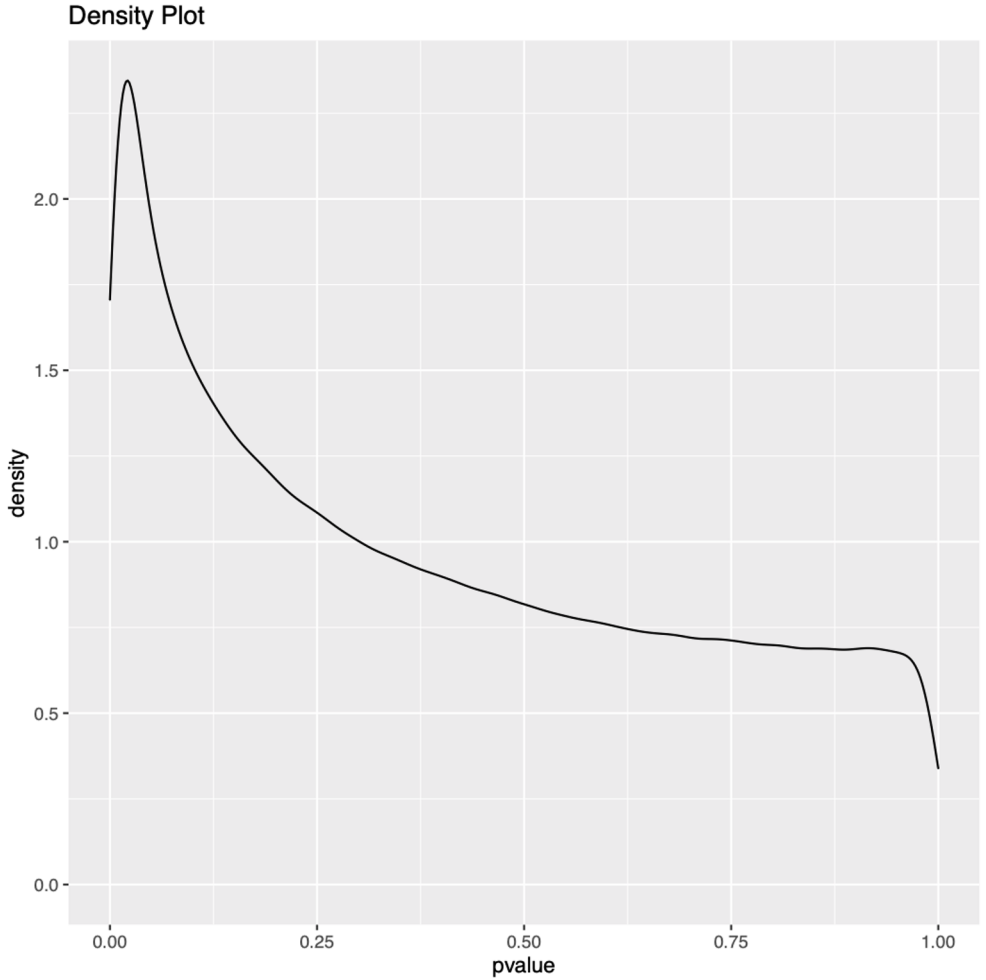
```
pvalHistogram4 <- pvalues %>%
  ggplot (aes(P.Value)) + geom_histogram()
```



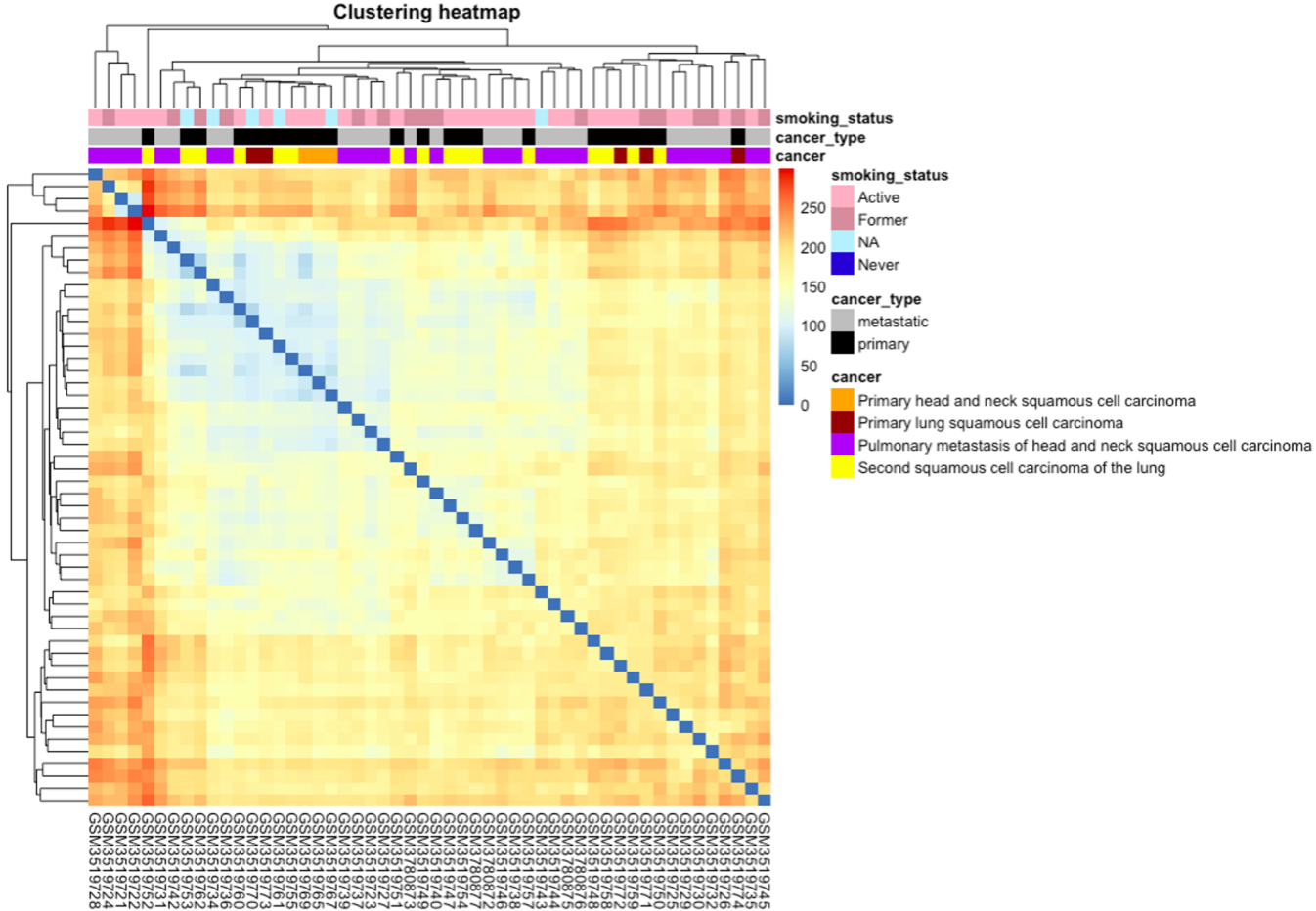
Quality Control

P-value Density Plot

```
data_subset %>%
  group_by(CG)%>%
  summarize( pvalue = t.test(value ~ cancer_type)$p.value) %>%
  ggplot(aes(x =pvalue)) + geom_density()+ggtitle("Density Plot")
```



Quality Control: Clustering Heatmap



Statistical Analysis: Methodologies used

Beta-values and M-values: summarize methylation and un-methylation intensities for each probe

Linear regression on M-values: the use of linear regression model to assess differential methylation in the context of multifactorial designed experiment (`~cancer_type*Age`), run on all probes

- Cancer type is primary vs metastatic

P-values: distribution of p-values across all tests to provide good diagnostics

T statistics: hypothesis testing to screen all possible hypotheses and find the ones that are statistically significant

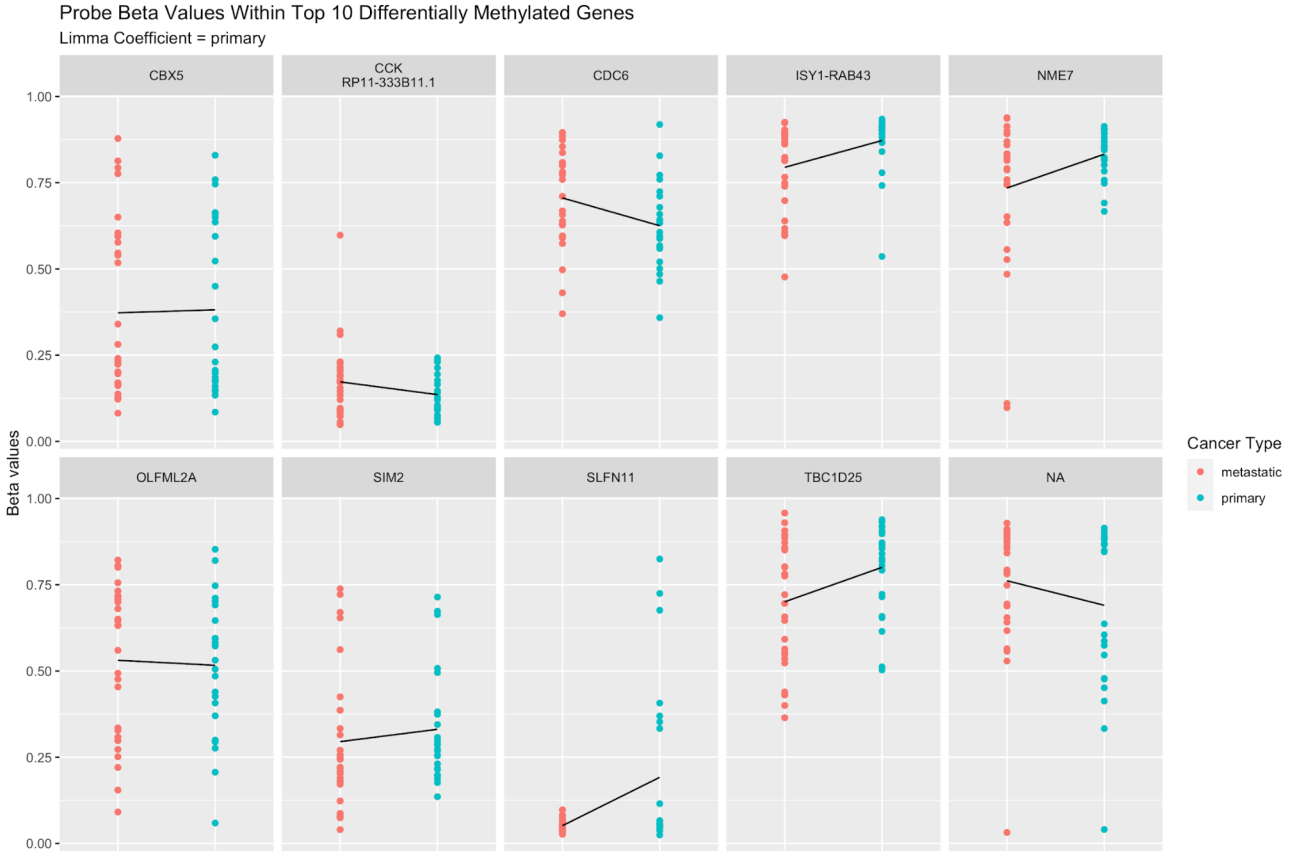
Limma Details

- Accounted for age as a covariate due to general effects of age on methylation
- Accounted for sex by only using male samples
- Final Limma model : Methylation~ Cancer_type+Age+ Age:Cancer_type
- Used the default Benjamini-Hochberg multiple test correction methods for the adjusted p values
- Could not use where the cancer origin was with cancer_type because we were missing one group (metastatic for lung) which caused problems in limma so decided to leave it out

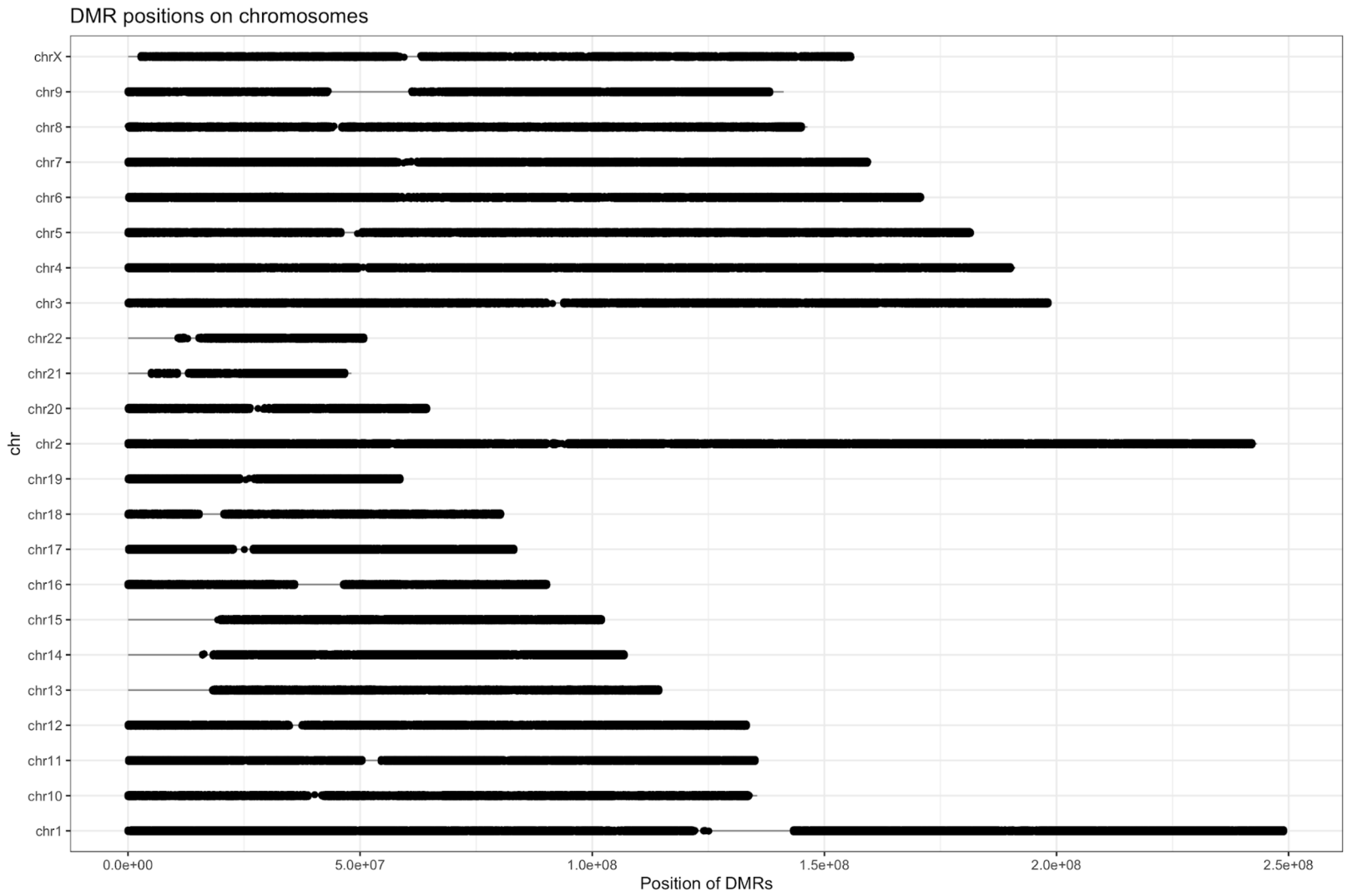
Top Table-Coefficient of Primary Cancer

Gene	ProbeID	Chr	logFC	T	P value	FDR
TBC1D25	cg14493612	X	-10.48	-4.88	1.03e-5	0.99
CDC6	cg21255171	17	-8.29	-4.72	1.79e-5	0.99
ISY1-RAB43	cg10716343	3	7.99	4.65	2.29e-5	0.99
NA	cg08216425	1	-14.17	-4.49	4.03e-5	0.99
CBX5	cg11713274	12	13.61	4.39	5.59e-5	0.99
NME7	cg04788627	1	9.84	4.38	5.73e-5	0.99
OLFML2A	cg04015541	9	10.76	4.15	1.25e-4	0.99
CCK	cg27100229	3	6.82	4.14	1.27e-4	0.99
SIM2	cg00937982	21	10.08	4.10	1.45e-4	0.99
SLFN11	cg18108623	17	10.71	4.00	2.02e-4	0.99

Results: Strip Plot



Results: Chromosome Plot



Pathway Analysis: KEGG

Gene Implicated	KEGG ID	Pathways
CBX5	ko05034	Alcoholism
CBX5	ko00310	Lysine degradation
CBX5	ko05322	Systemic lupus erythematosus
CBX5	ko05202	Transcriptional misregulation in cancer
CBX5	ko05203	Viral carcinogenesis
CCK	ko04974	Protein digestion and absorption
CDC6	ko05219	Bladder cancer
CDC6	ko05224	Breast cancer

Discussion

- Adjusted P values showed no significant genes that are differentially methylated in the primary compared to metastatic
- There was no significant differential methylation by age or age:cancer_type interaction
- Due to unbalanced design (low number of samples), we decided not to separate the primary cancer into head and neck or lung cancer
- Overall sample size was still small so it would be interesting to look at larger numbers of primary lung and metastatic(pulmonary) HNSC

Challenges

- Large dataset
  - Computationally expensive for statistical analysis and for plotting, leading to RStudio crashing and frozen computers
- Unbalanced sample sizes
  - By cancer type : 28 metastatic vs 24 primary
  - By cancer: 3 primary HNSC vs 3 Primary lung squamous cell carcinoma vs 16 Primary Second Squamous Cell Lung Carcinoma vs 28 Metastatic HNSC
  - By cancer location: 31 HNSC vs 19 Lung
- Inconsistent metadata collection
  - Some samples were missing information for smoking, e.g. some smokers had # packs per day while others did not
- Incompatible R packages
  - ErmineR caused issues with various versions of Java (even Paul couldn’t figure it out)

Conclusions

We fail to reject the null hypothesis. However, due to the small sample sizes and unbalanced design, we lack sufficient evidence to definitively conclude that there is no differential methylation between primary (HNSC/LUSC) and metastatic cancers (HNSC).

Take Home Messages:

- Select datasets with balanced metadata and a *larger than enough* sample size to maximize statistical power
- Select size of datasets based on available computational resources
- If given raw data and processed data, use the raw data and do all preprocessing steps (i.e. normalization) yourself to ensure quality analysis
- GitHub is great for collaborative high-dimensional biology analyses!