Lung cancer dimensionality reduction

load R packages

```
library(tidyverse)
library(umap)
```

load gene expression data in FPKM format

```
luad_fpkm <- read_tsv("./data/processed/TCGA-GDC/rna_seq/tcga_luad_PrimaryTumor_FPKM.txt.gz")
lusc_fpkm <- read_tsv("./data/processed/TCGA-GDC/rna_seq/tcga_lusc_PrimaryTumor_FPKM.txt.gz")
```

load samples metadata

```
luad_metadata <- read_tsv("./data/processed/metadata/tcga_luad_metadata_pancancer_atlas.txt")
lusc_metadata <- read_tsv("./data/processed/metadata/tcga_lusc_metadata_pancancer_atlas.txt")
```

load gene annotation

```
gene_annot <- read_tsv("./data/raw/TCGA-GDC/rna_seq/gene_annotation/gencode.gene.info.v22.tsv")
```

select only protein coding and lincRNA genes

```
sel_genes <- gene_annot %>%
  filter(gene_type %in% c("protein_coding", "lincRNA"))

luad_fpkm <- luad_fpkm %>%
  semi_join(sel_genes, by = c("gene" = "gene_id"))

lusc_fpkm <- lusc_fpkm %>%
  semi_join(sel_genes, by = c("gene" = "gene_id"))
```

select only the genes with median fpkm > 1 and transform values to log2

```
luad_fpkm <- luad_fpkm %>%
  pivot_longer(-gene, names_to = "sample", values_to = "fpkm") %>%
  group_by(gene) %>%
  filter(median(fpkm) > 1) %>%
  ungroup() %>%
  mutate(fpkm = log2(fpkm + 1)) %>%
  mutate(sample = str_sub(sample, 1, 12)) %>%
  pivot_wider(names_from = "sample", values_from = "fpkm")

lusc_fpkm <- lusc_fpkm %>%
  pivot_longer(-gene, names_to = "sample", values_to = "fpkm") %>%
  group_by(gene) %>%
  filter(median(fpkm) > 1) %>%
  ungroup() %>%
  mutate(fpkm = log2(fpkm + 1)) %>%
  mutate(sample = str_sub(sample, 1, 12)) %>%
  pivot_wider(names_from = "sample", values_from = "fpkm")
```

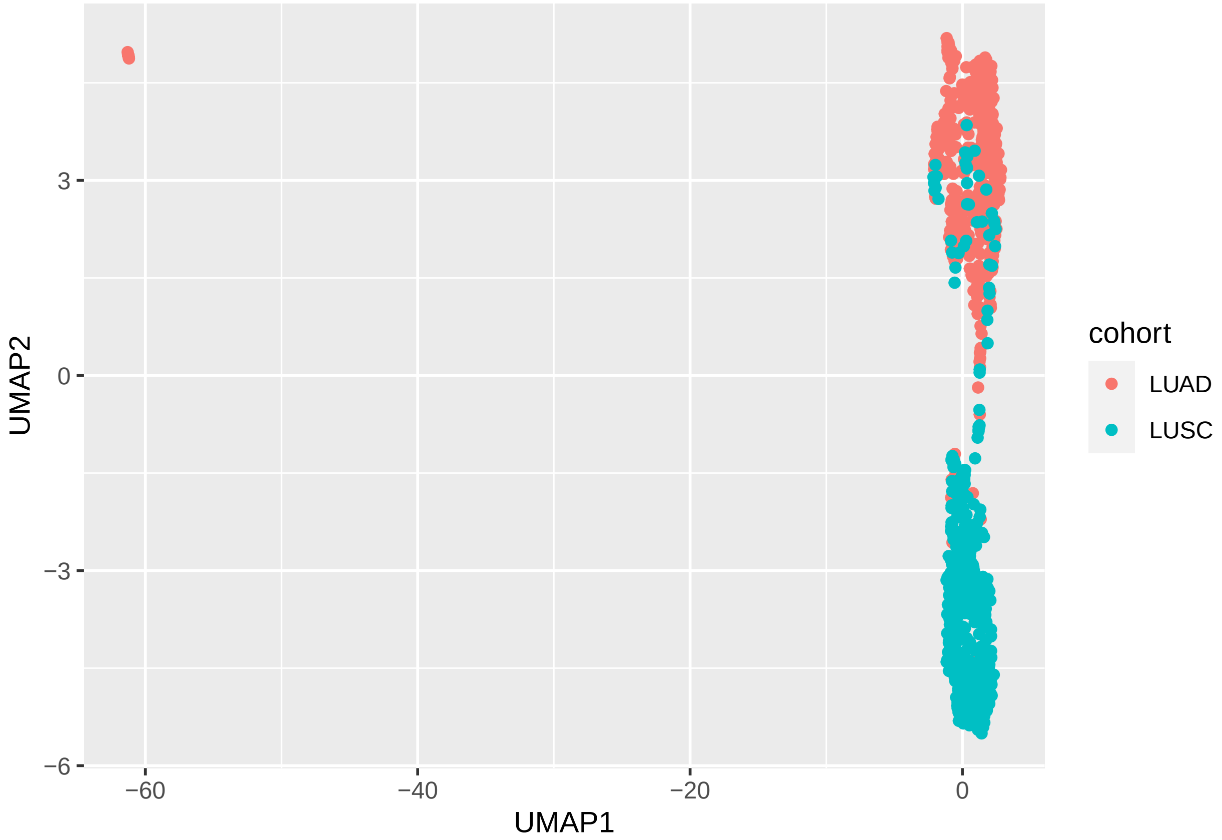
dimensionality reduction across both cohorts

set up a matrix with both cohorts

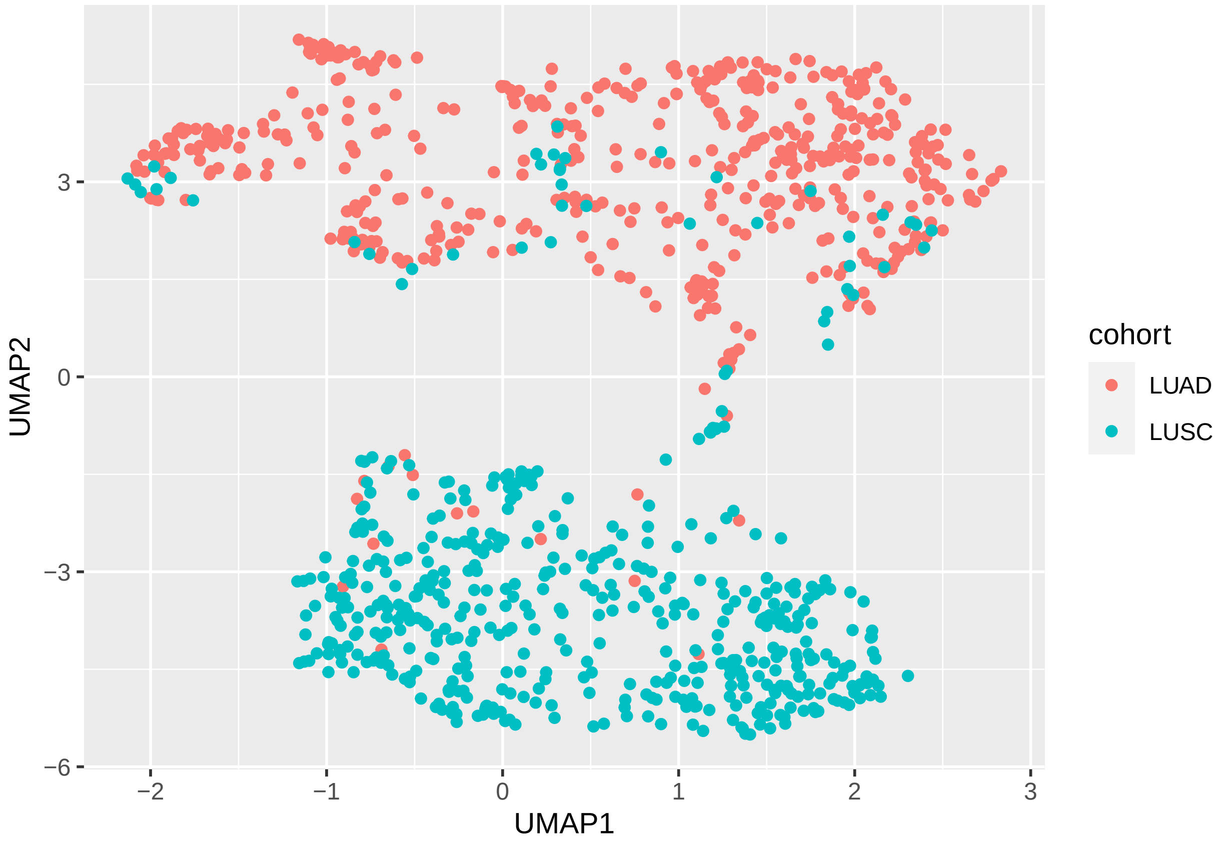
```
tcga_lung <- inner_join(luad_fpkm, lusc_fpkm, by = "gene") %>%
  column_to_rownames(var = "gene") %>%
  as.matrix() %>%
  t()
```

UMAP analysis across both cohorts

```
tcga_lung_umap <- umap(tcga_lung)

tcga_metadata <- bind_rows(luad_metadata[, c("sample", "tcga_cohort")], lusc_metadata[, c("sample", "tc
  rename(cohort = tcga_cohort)

tcga_lung_umap_plot <- tcga_lung_umap %>%
  pluck("layout") %>%
  as.data.frame() %>%
  rownames_to_column(var = "sample") %>%
  as_tibble() %>%
  rename(UMAP1 = V1, UMAP2 = V2) %>%
  inner_join(tcga_metadata, by = "sample") %>%
  ggplot(mapping = aes(x = UMAP1, y = UMAP2, color = cohort)) +
    geom_point()
tcga_lung_umap_plot
```

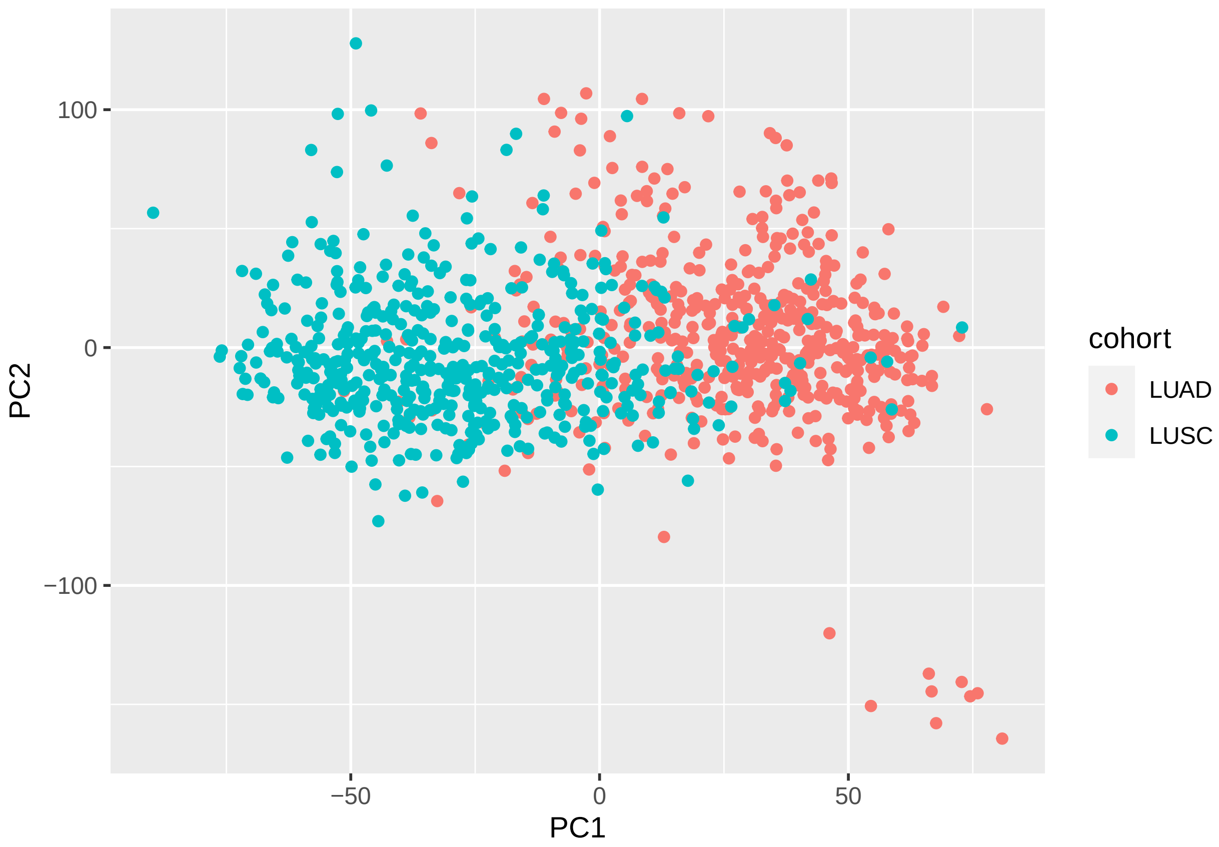


```r
tcga_lung_umap_plot <- tcga_lung_umap %>%
  pluck("layout") %>%
  as.data.frame() %>%
  rownames_to_column(var = "sample") %>%
  as_tibble() %>%
  rename(UMAP1 = V1, UMAP2 = V2) %>%
  filter(UMAP1 > -20) %>%
  inner_join(tcga_metadata, by = "sample") %>%
  ggplot(mapping = aes(x = UMAP1, y = UMAP2, color = cohort)) +
    geom_point()
tcga_lung_umap_plot
```



PCA analysis across both cohorts

```r
tcga_lung_pca <- prcomp(tcga_lung, center = T, scale. = T)

tcga_lung_pca_plot <- tcga_lung_pca %>%
  pluck("x") %>%
  as.data.frame() %>%
  rownames_to_column(var = "sample") %>%
  as_tibble() %>%
  inner_join(tcga_metadata, by = "sample") %>%
  ggplot(mapping = aes(x = PC1, y = PC2, color = cohort)) +
  geom_point()
tcga_lung_pca_plot
```



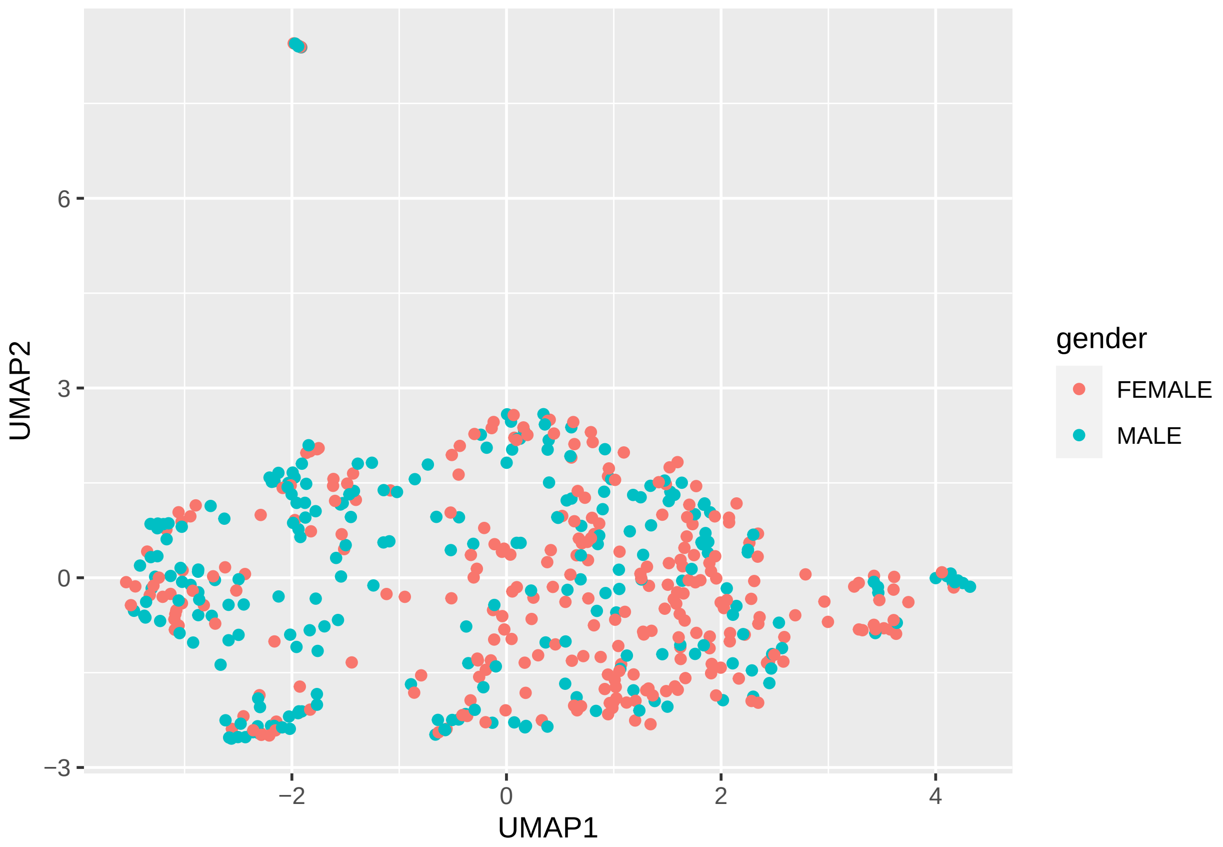
dimensionality reduction across LUAD cohort

set up a matrix with LUAD cohort

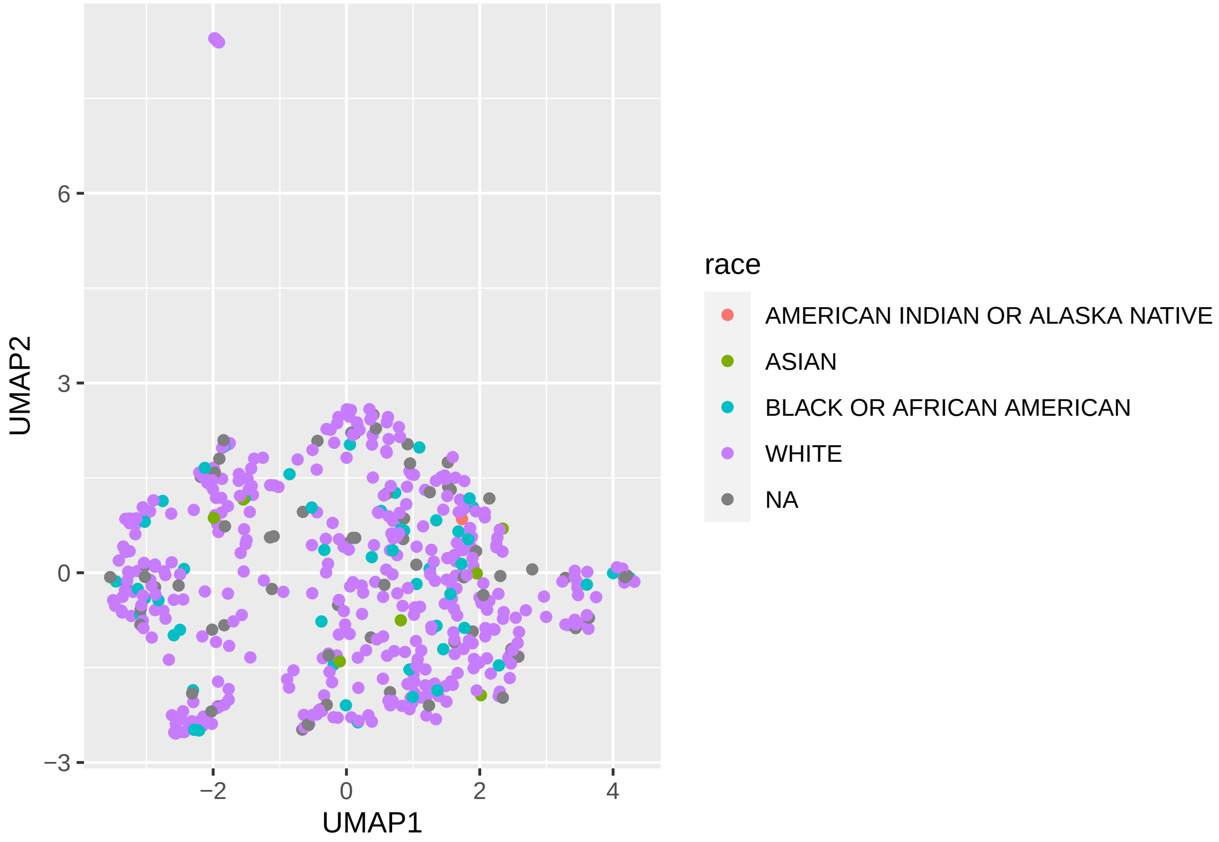
```
luad_mat <- luad_fpkm %>%
  column_to_rownames(var = "gene") %>%
  as.matrix() %>%
  t()
```

UMAP analysis across LUAD cohort

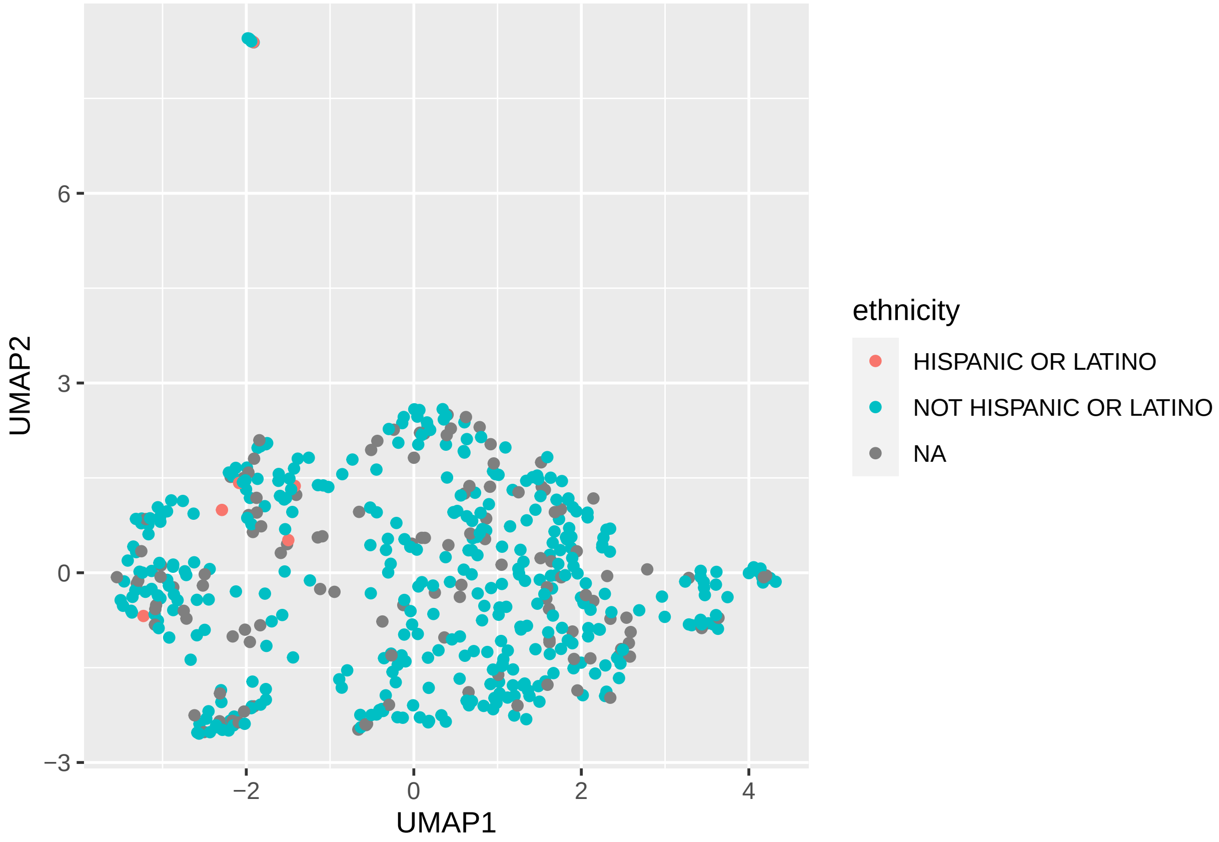
```
luad_umap <- umap(luad_mat)

luad_umap_temp_plot <- luad_umap %>%
  pluck("layout") %>%
  as.data.frame() %>%
  rownames_to_column(var = "sample") %>%
  as_tibble() %>%
  rename(UMAP1 = V1, UMAP2 = V2) %>%
  inner_join(luad_metadata, by = "sample") %>%
  ggplot(mapping = aes(x = UMAP1, y = UMAP2))

luad_umap_plot <- luad_umap_temp_plot +
  geom_point(mapping = aes(color = gender))
luad_umap_plot
```



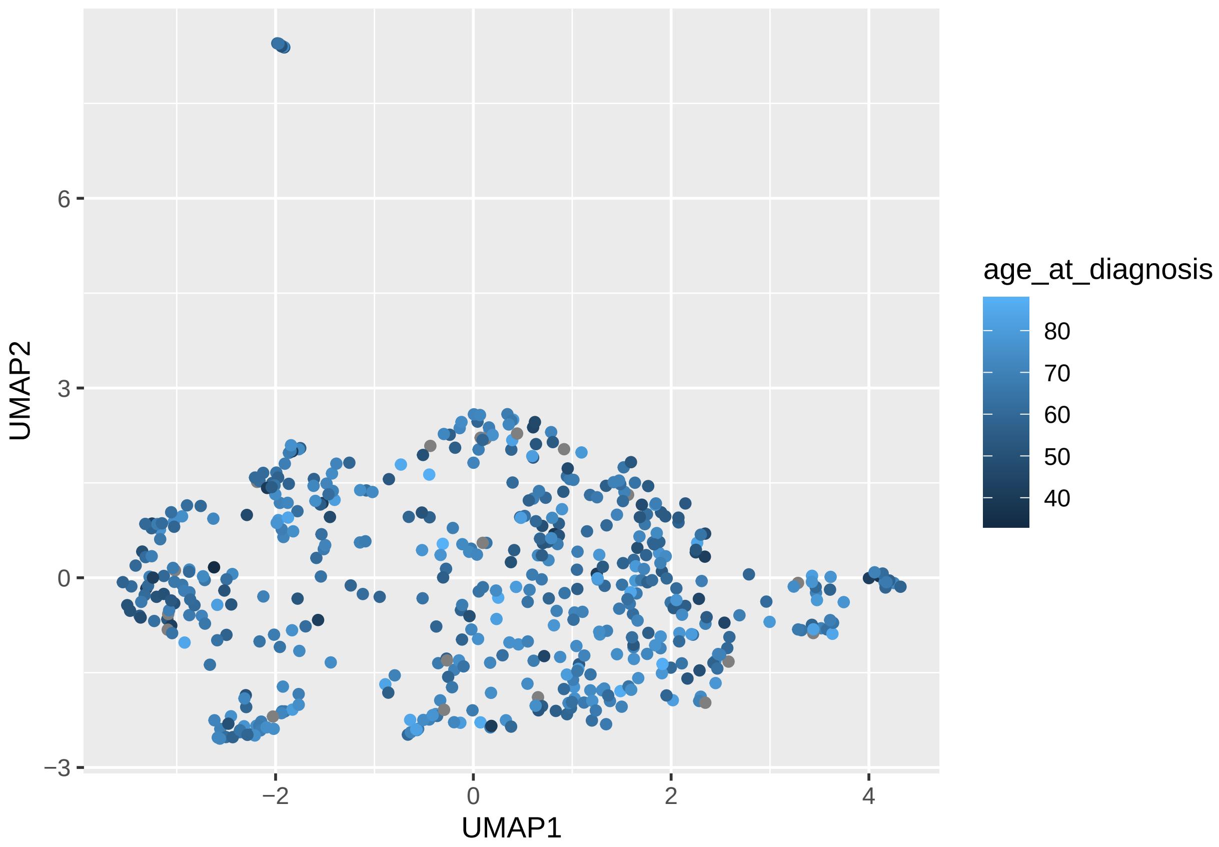
```
luad_umap_plot <- luad_umap_temp_plot +
  geom_point(mapping = aes(color = race))
luad_umap_plot
```



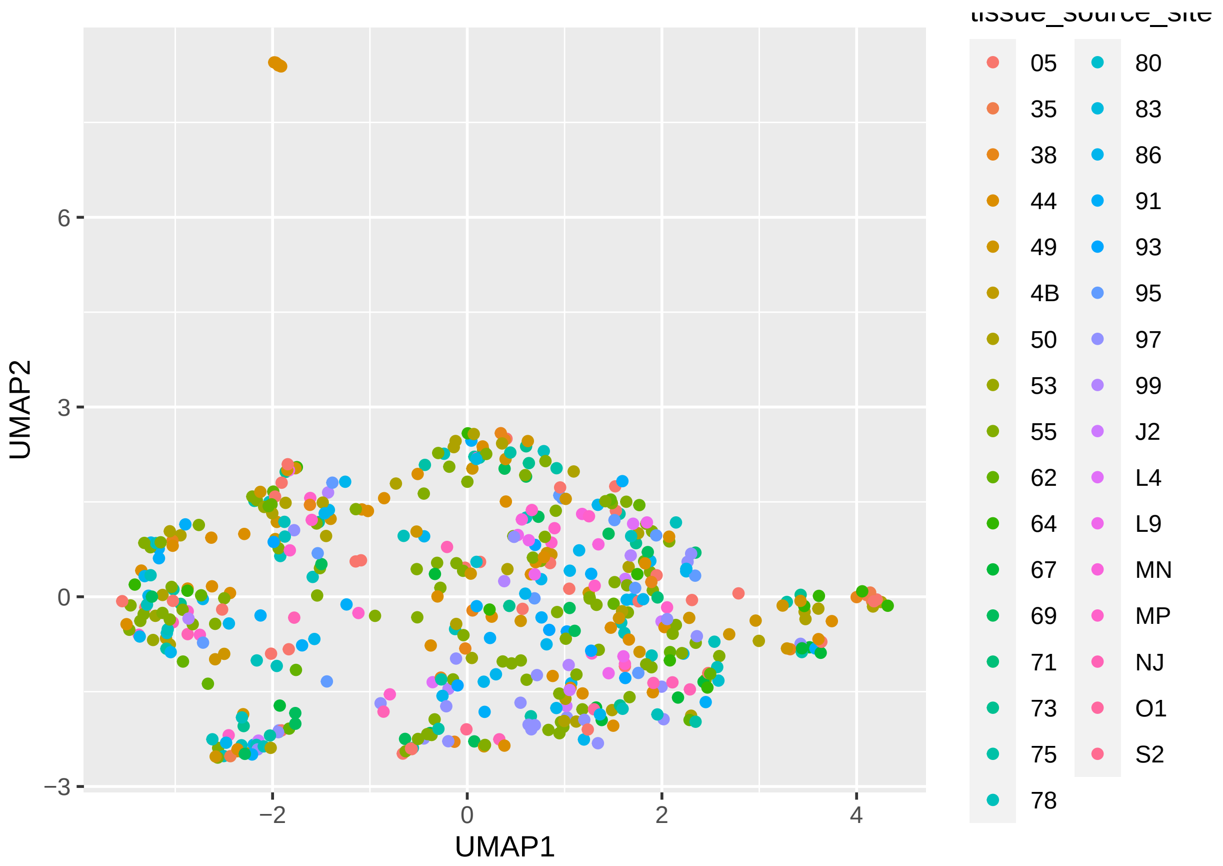
```
luad_umap_plot <- luad_umap_temp_plot +
  geom_point(mapping = aes(color = ethnicity))
luad_umap_plot
```



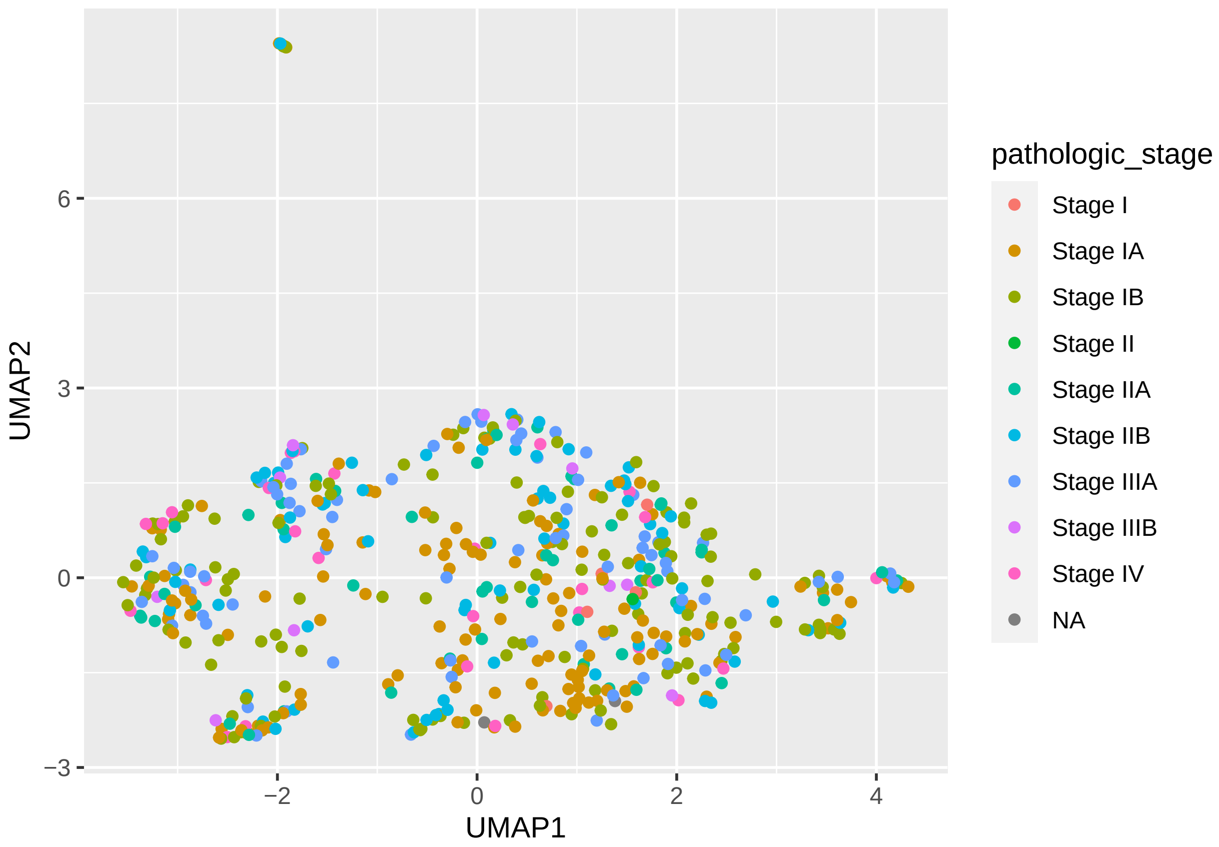
```
luad_umap_plot <- luad_umap_temp_plot +
  geom_point(mapping = aes(color = age_at_diagnosis))
luad_umap_plot
```



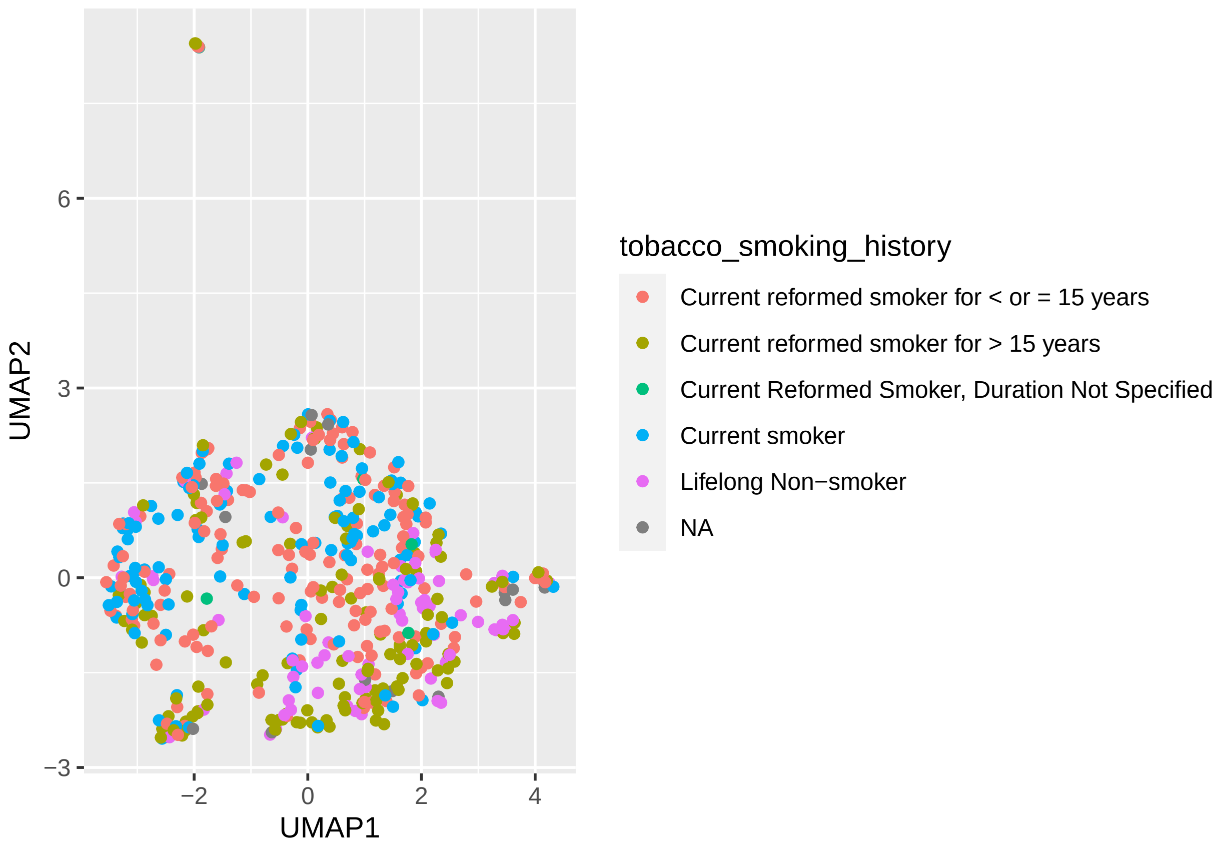
```
luad_umap_plot <- luad_umap_temp_plot +
  geom_point(mapping = aes(color = tissue_source_site))
luad_umap_plot
```



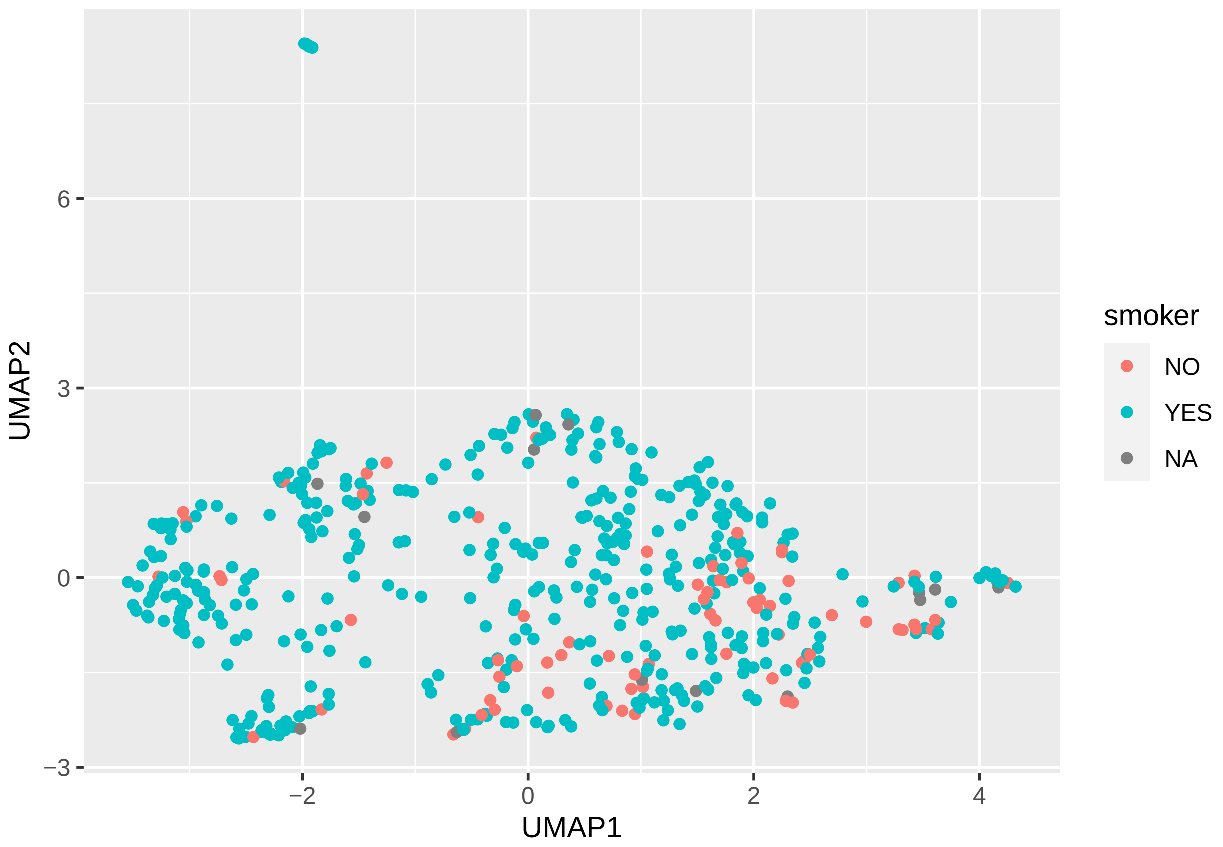
```
luad_umap_plot <- luad_umap_temp_plot +
  geom_point(mapping = aes(color = pathologic_stage))
luad_umap_plot
```



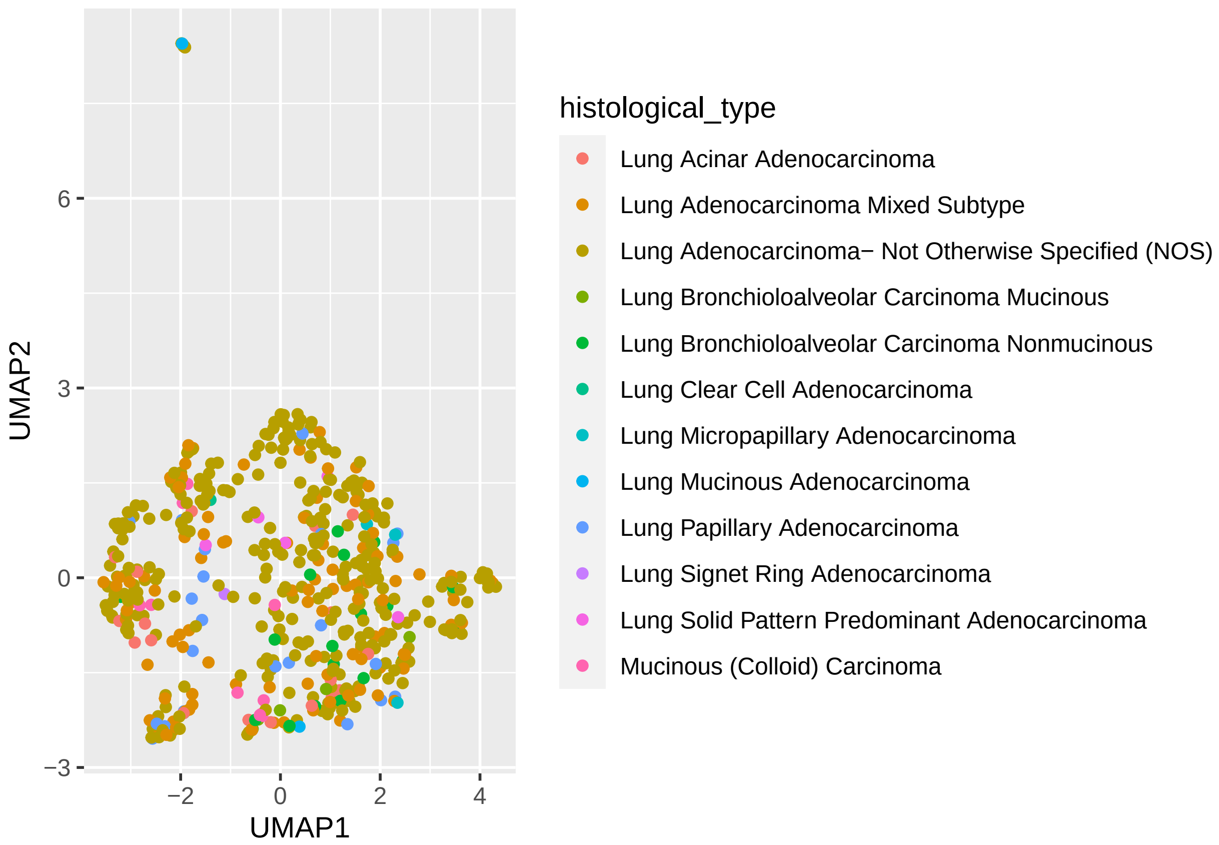
```
luad_umap_plot <- luad_umap_temp_plot +
  geom_point(mapping = aes(color = tobacco_smoking_history))
luad_umap_plot
```



```
luad_umap_plot <- luad_umap_temp_plot +
  geom_point(mapping = aes(color = smoker))
luad_umap_plot
```

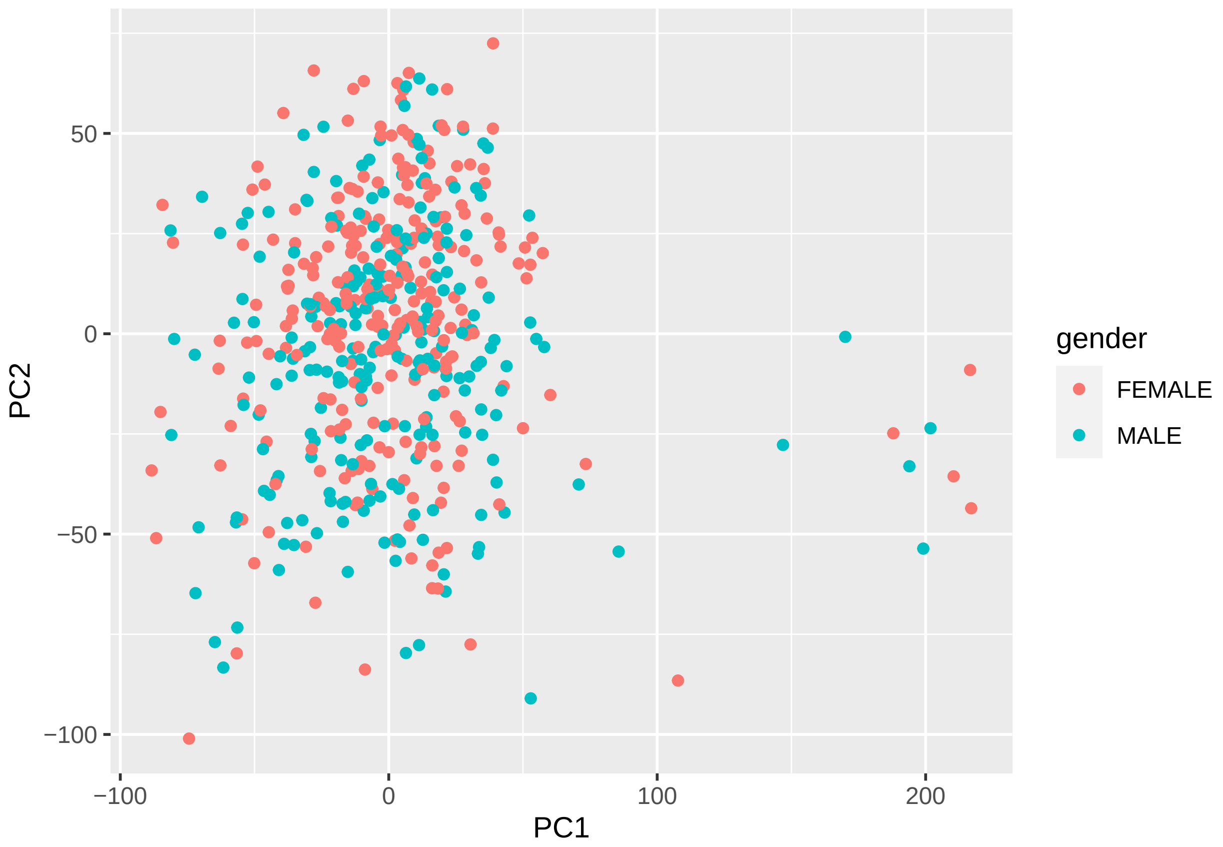


```
luad_umap_plot <- luad_umap_temp_plot +
  geom_point(mapping = aes(color = histological_type))
luad_umap_plot
```

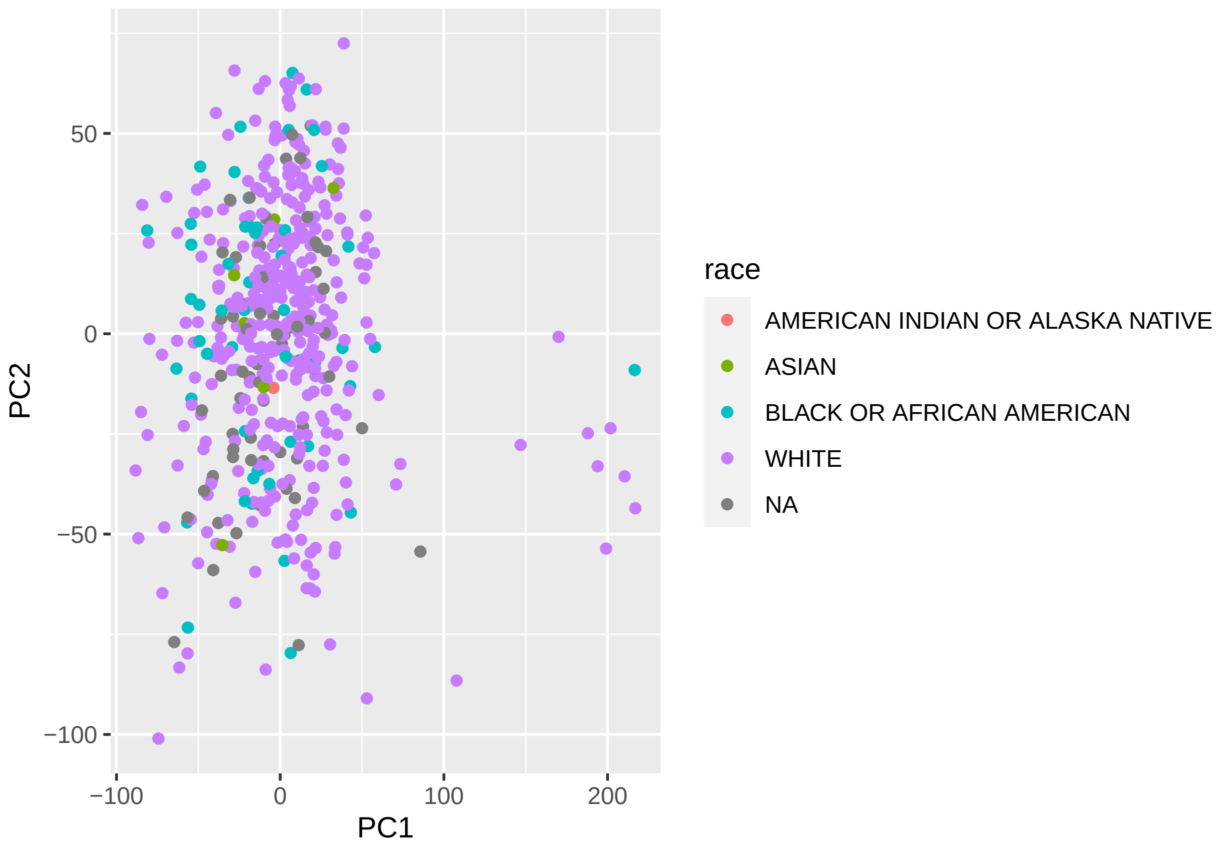


PCA analysis across LUAD cohort

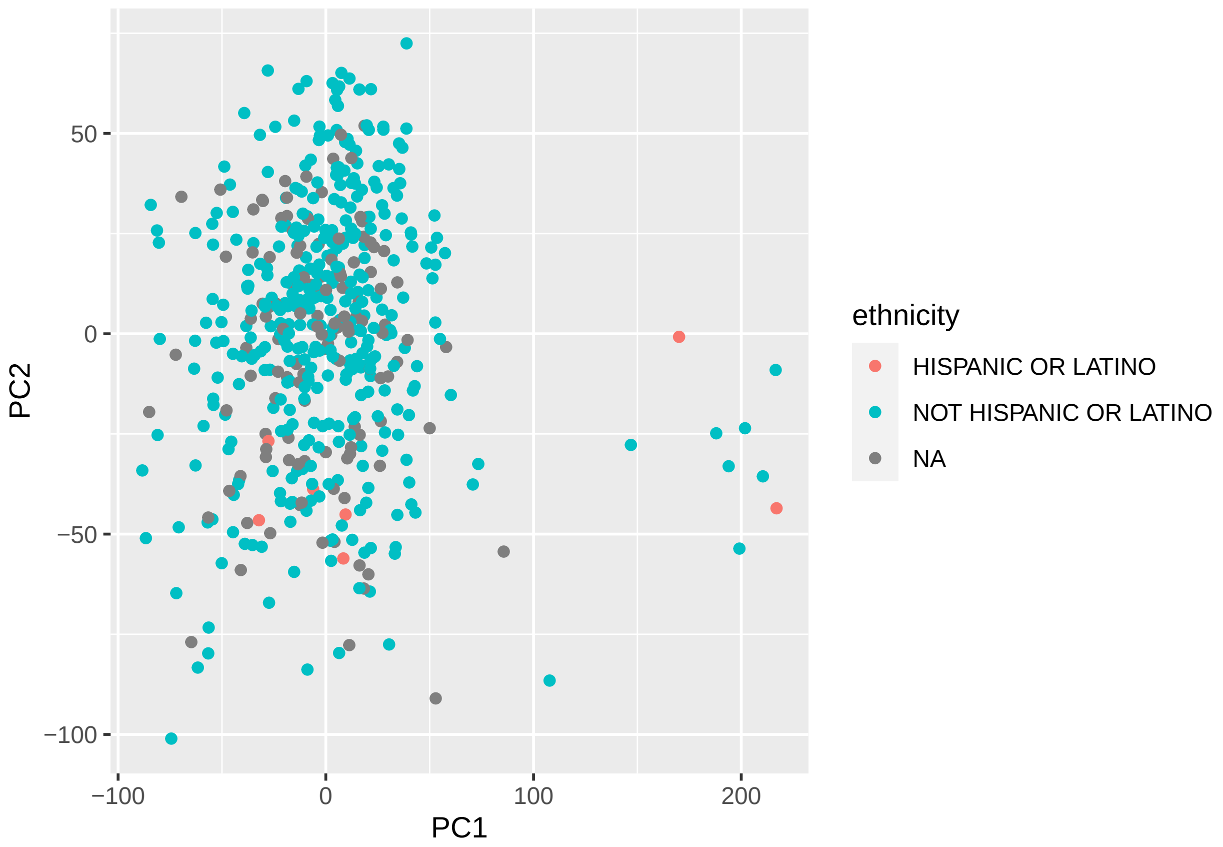
```
luad_pca <- prcomp(luad_mat, center = T, scale. = T)

luad_pca_temp_plot <- luad_pca %>%
  pluck("x") %>%
  as.data.frame() %>%
  rownames_to_column(var = "sample") %>%
  as_tibble() %>%
  inner_join(luad_metadata, by = "sample") %>%
  ggplot(mapping = aes(x = PC1, y = PC2))

luad_pca_plot <- luad_pca_temp_plot +
  geom_point(mapping = aes(color = gender))
luad_pca_plot
```



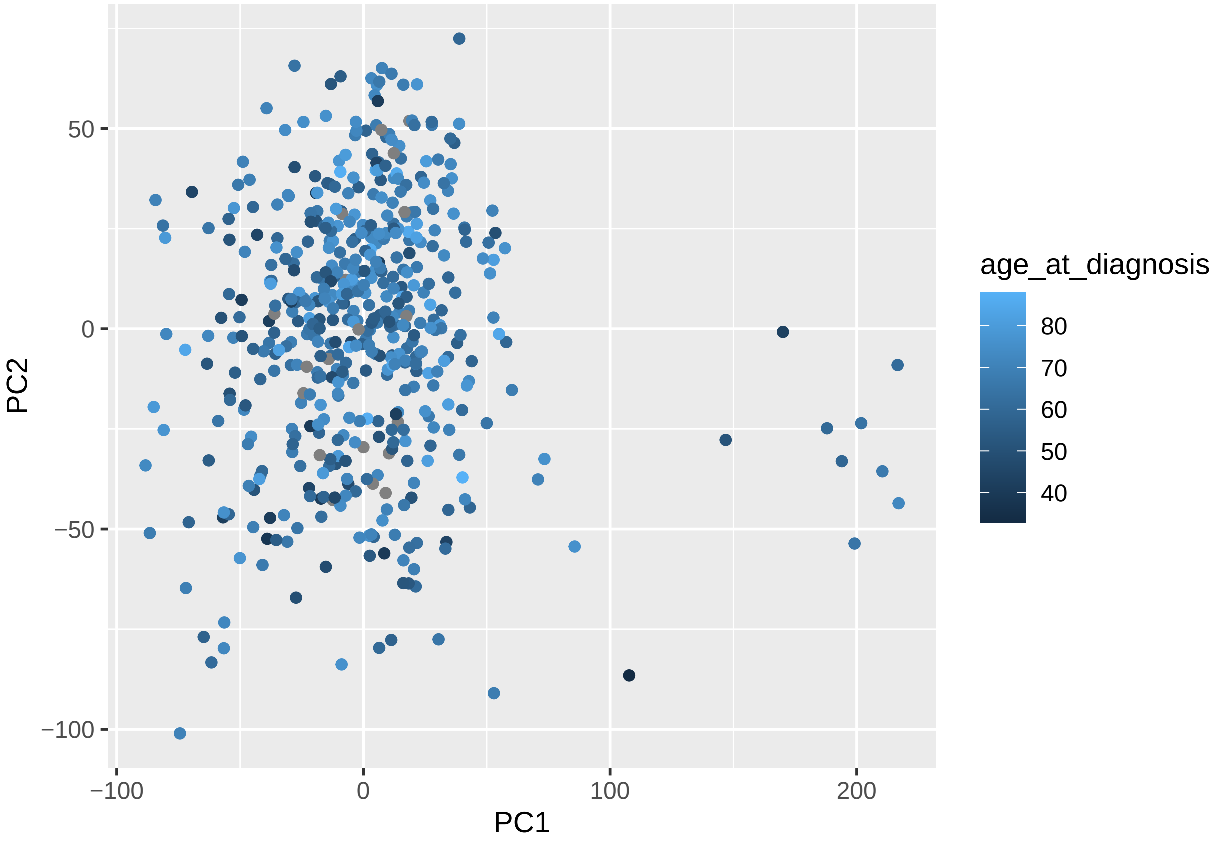
```
luad_pca_plot <- luad_pca_temp_plot +
  geom_point(mapping = aes(color = race))
luad_pca_plot
```



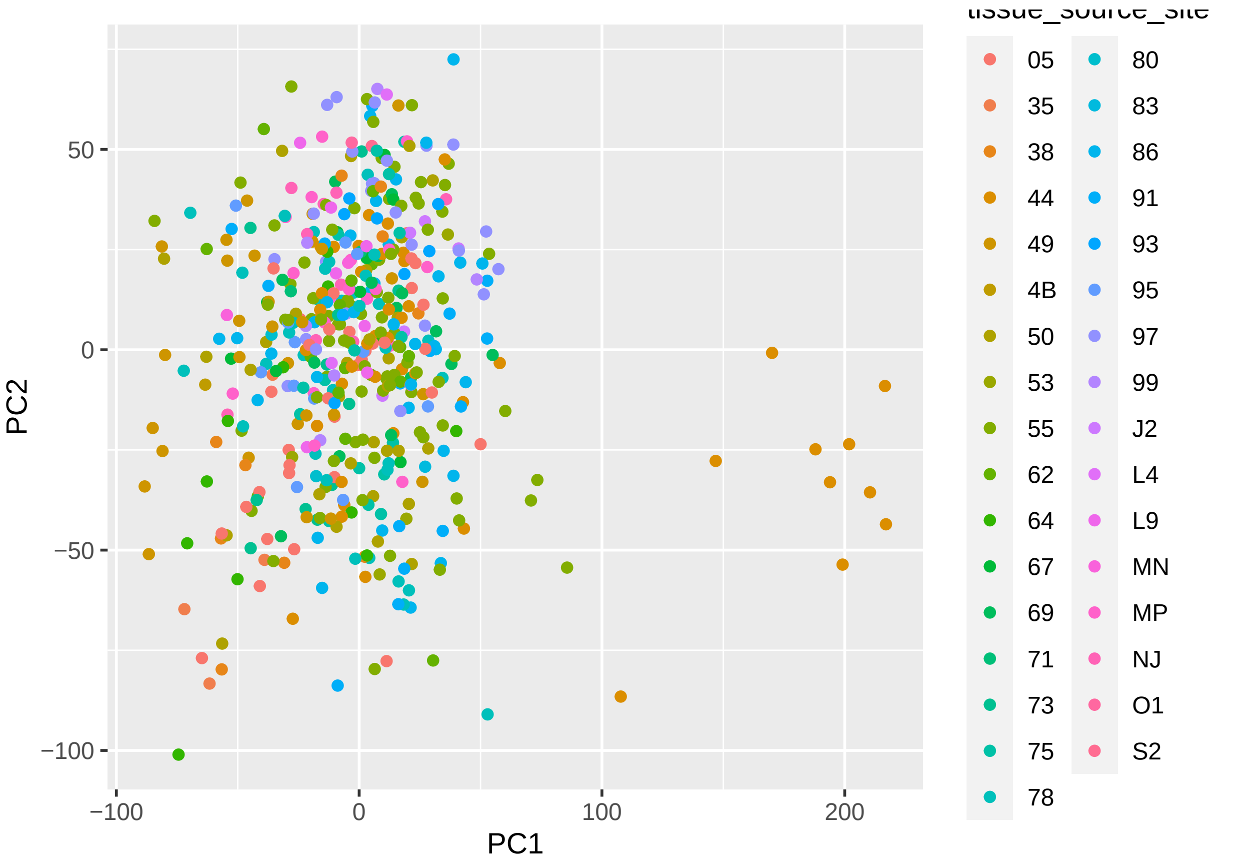
```
luad_pca_plot <- luad_pca_temp_plot +
  geom_point(mapping = aes(color = ethnicity))
luad_pca_plot
```



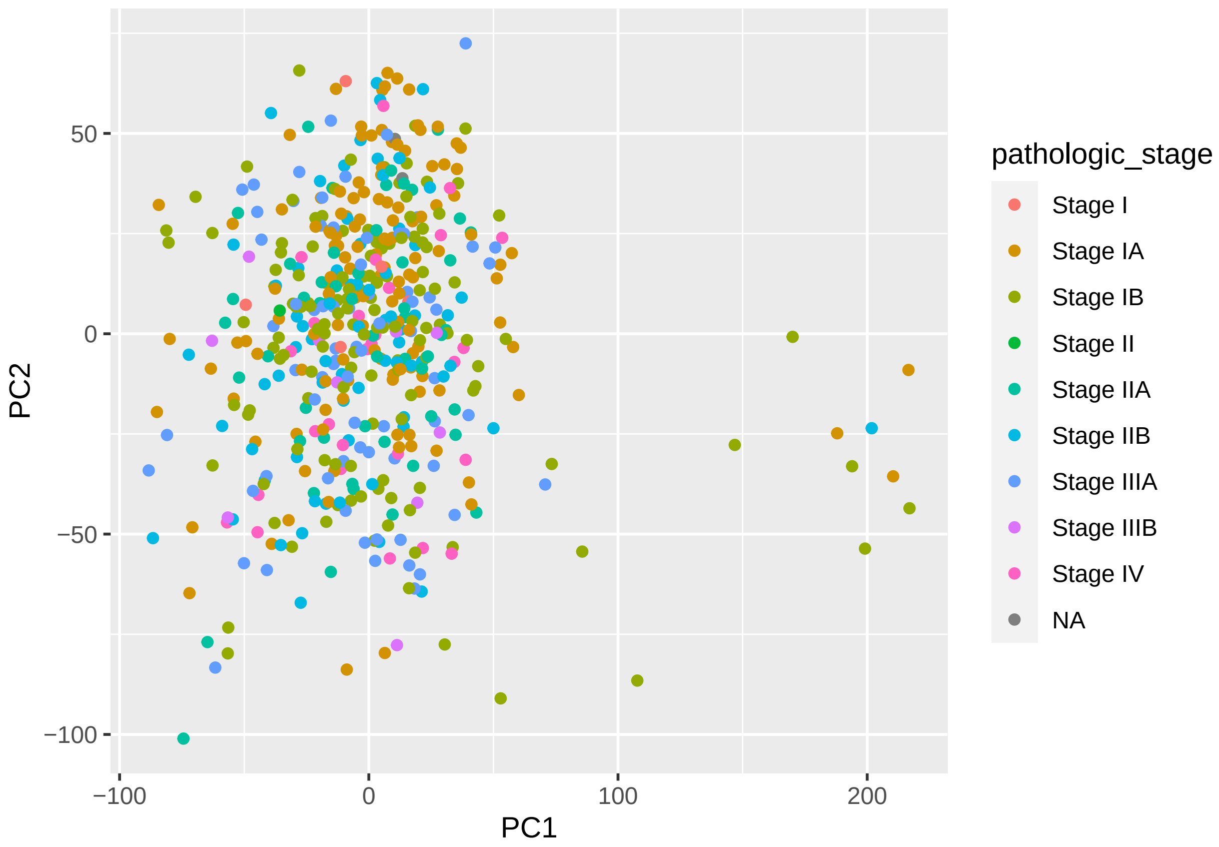
```
luad_pca_plot <- luad_pca_temp_plot +
  geom_point(mapping = aes(color = age_at_diagnosis))
luad_pca_plot
```



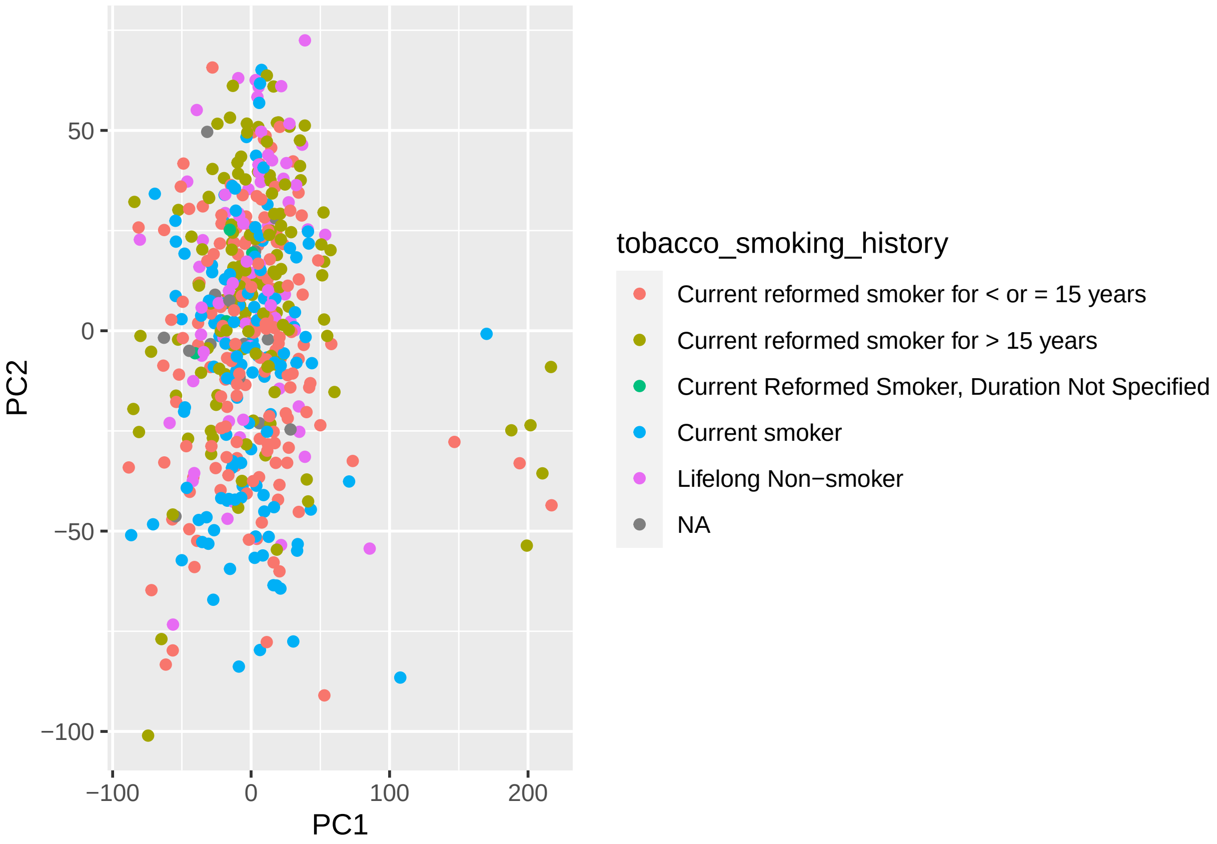
```
luad_pca_plot <- luad_pca_temp_plot +
  geom_point(mapping = aes(color = tissue_source_site))
luad_pca_plot
```



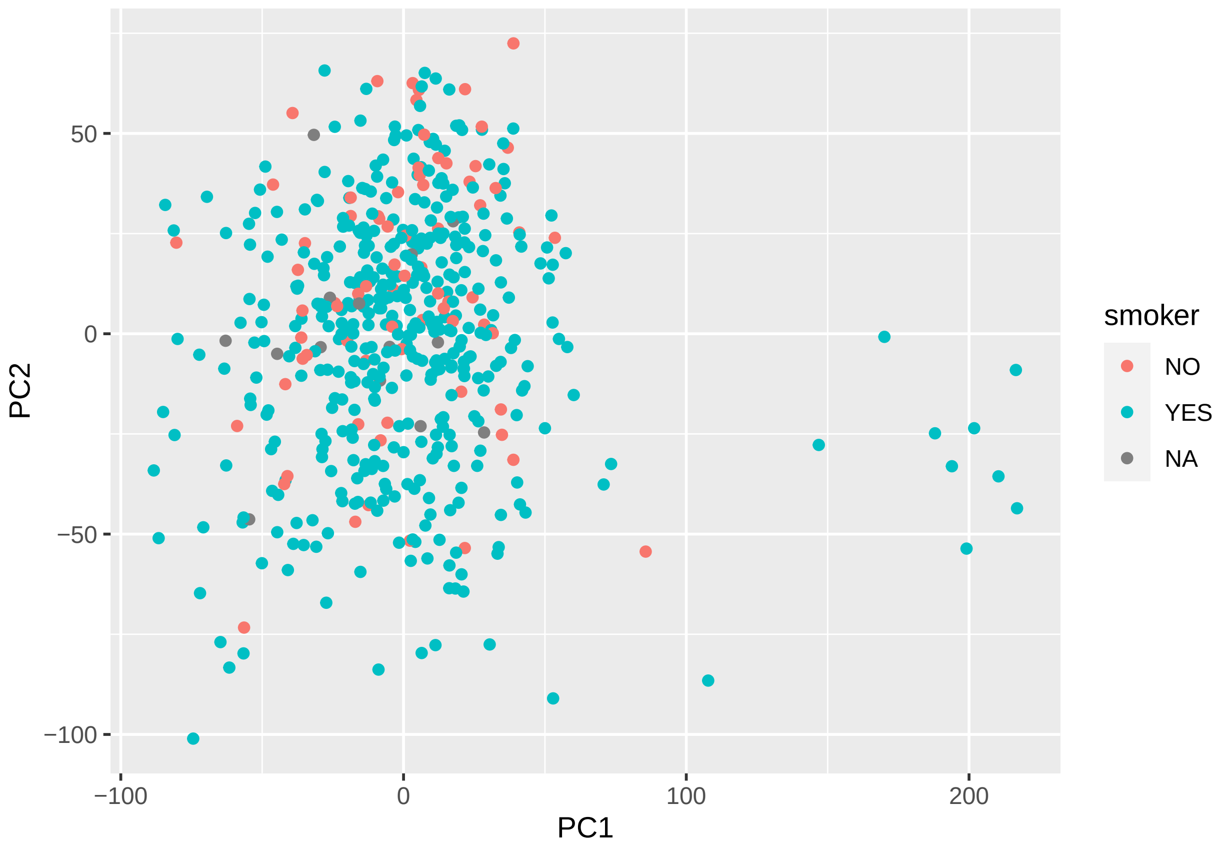
```
luad_pca_plot <- luad_pca_temp_plot +
  geom_point(mapping = aes(color = pathologic_stage))
luad_pca_plot
```



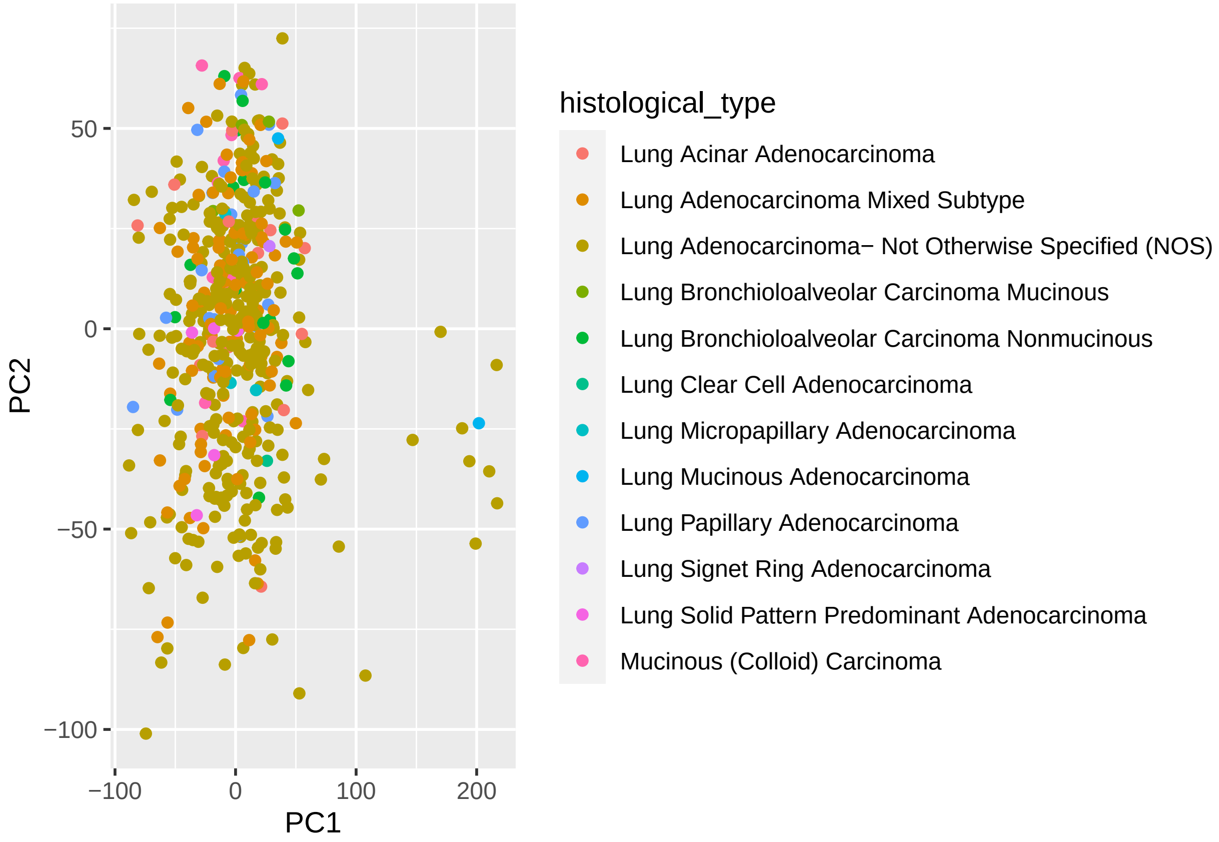
```
luad_pca_plot <- luad_pca_temp_plot +
  geom_point(mapping = aes(color = tobacco_smoking_history))
luad_pca_plot
```



```
luad_pca_plot <- luad_pca_temp_plot +
  geom_point(mapping = aes(color = smoker))
luad_pca_plot
```

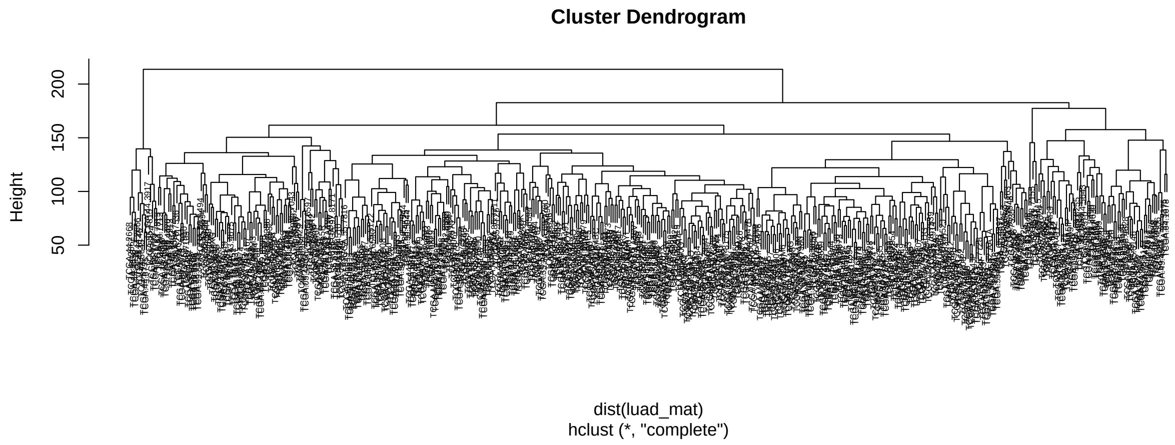


```
luad_pca_plot <- luad_pca_temp_plot +
  geom_point(mapping = aes(color = histological_type))
luad_pca_plot
```



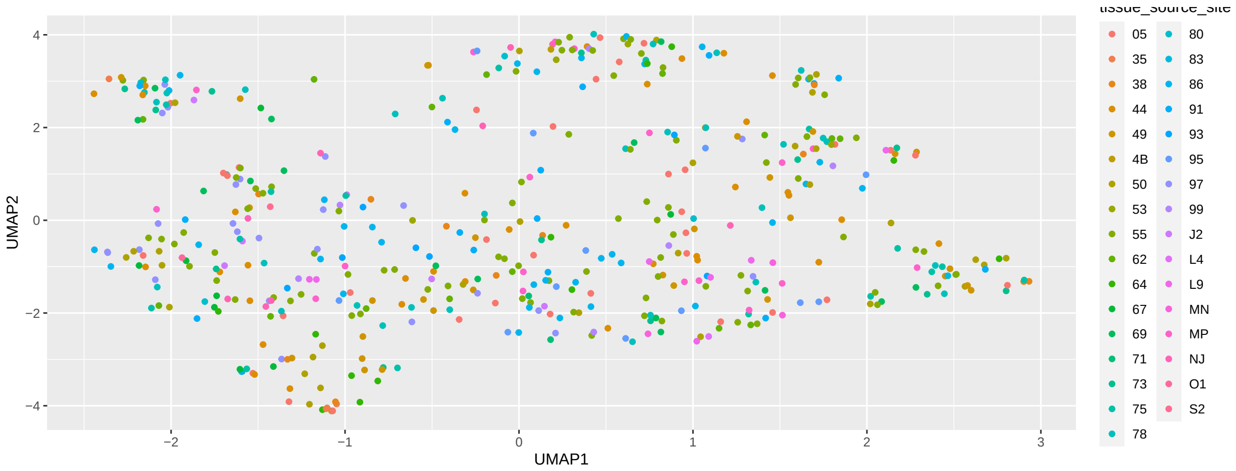
hierachical clustering to select outlier samples use default distance measure (eucledian) and agglomeration method (complete)

```
luad_hc <- hclust(dist(luad_mat))

plot(luad_hc, cex = 0.5)
```

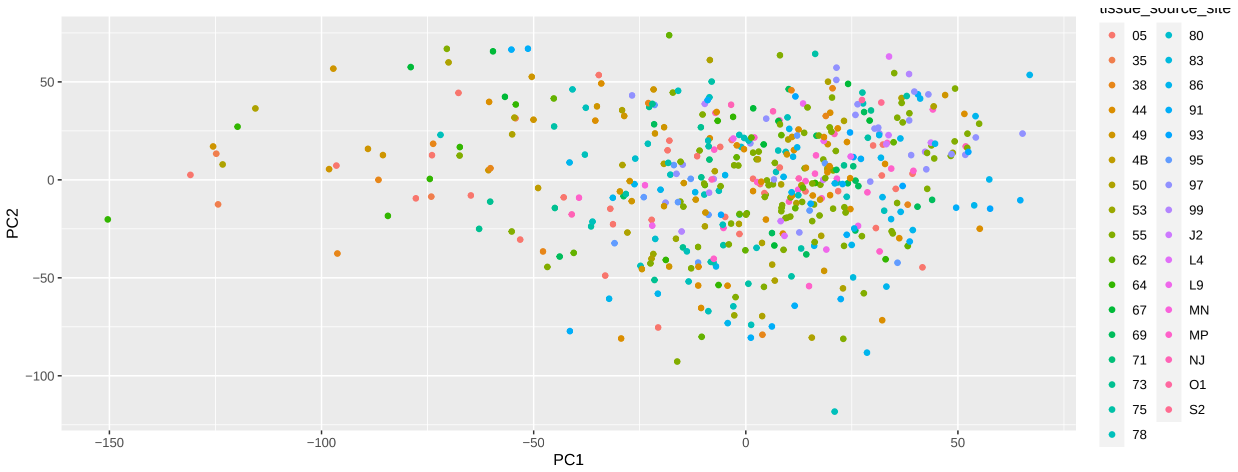


```
luad_clusters <- cutree(luad_hc, h = 200)
luad_clusters <- tibble(sample = names(luad_clusters), cluster = unname(luad_clusters))

write_tsv(x = luad_clusters, file = "./output/files/02_exploratory_analysis/tcga_luad_outlier_samples_h
```

```r
luad_umap2 <- umap(luad_mat[rownames(luad_mat) %in% luad_clusters[luad_clusters$cluster == 1, ]$sample,

luad_umap_temp_plot2 <- luad_umap2 %>%
  pluck("layout") %>%
  as.data.frame() %>%
  rownames_to_column(var = "sample") %>%
  as_tibble() %>%
  rename(UMAP1 = V1, UMAP2 = V2) %>%
  inner_join(luad_metadata, by = "sample") %>%
  ggplot(mapping = aes(x = UMAP1, y = UMAP2))

luad_umap_plot2 <- luad_umap_temp_plot2 +
  geom_point(mapping = aes(color = tissue_source_site))
luad_umap_plot2
```



```r
luad_pca2 <- prcomp(luad_mat[rownames(luad_mat) %in% luad_clusters[luad_clusters$cluster == 1, ]$sample

luad_pca_temp_plot2 <- luad_pca2 %>%
  pluck("x") %>%
  as.data.frame() %>%
  rownames_to_column(var = "sample") %>%
  as_tibble() %>%
  inner_join(luad_metadata, by = "sample") %>%
  ggplot(mapping = aes(x = PC1, y = PC2))

luad_pca_plot2 <- luad_pca_temp_plot2 +
  geom_point(mapping = aes(color = tissue_source_site))
luad_pca_plot2
```



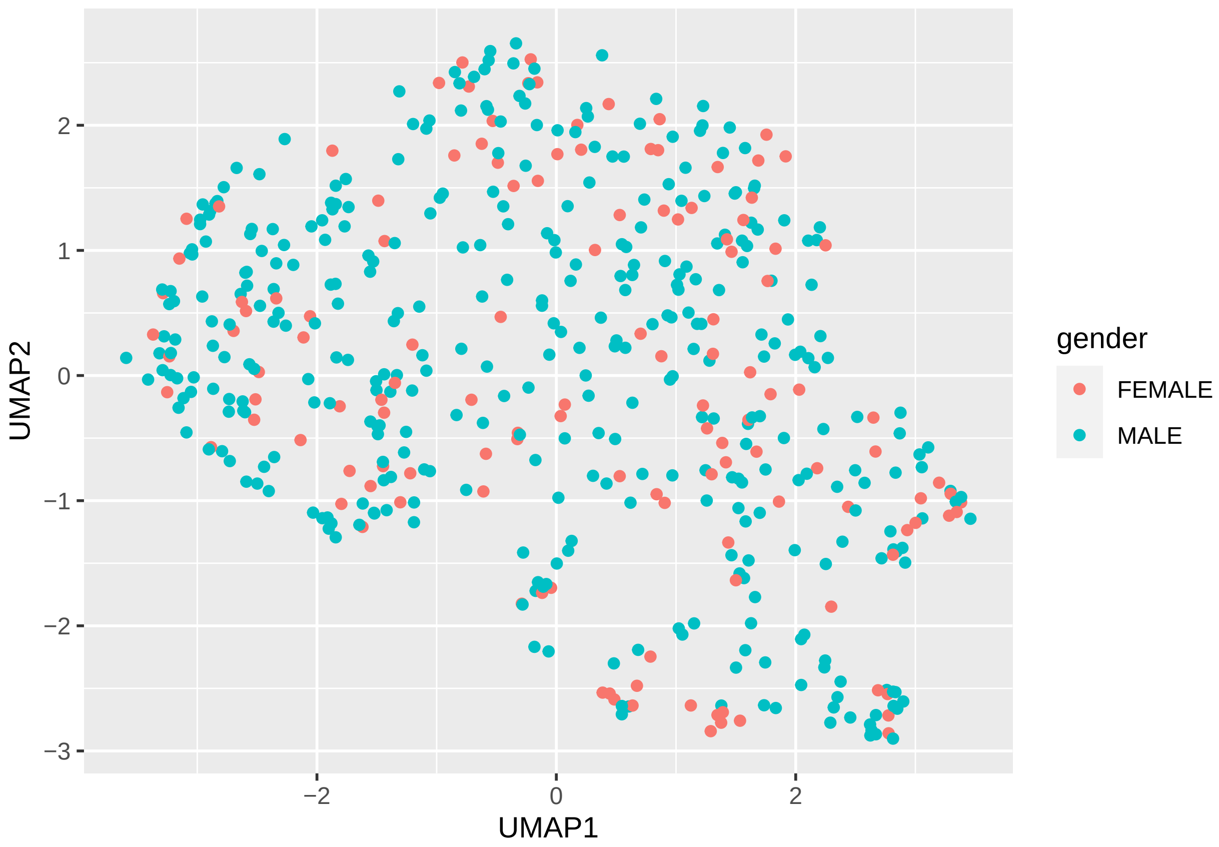
dimensionality reduction across LUSC cohort

set up a matrix with LUSC cohort

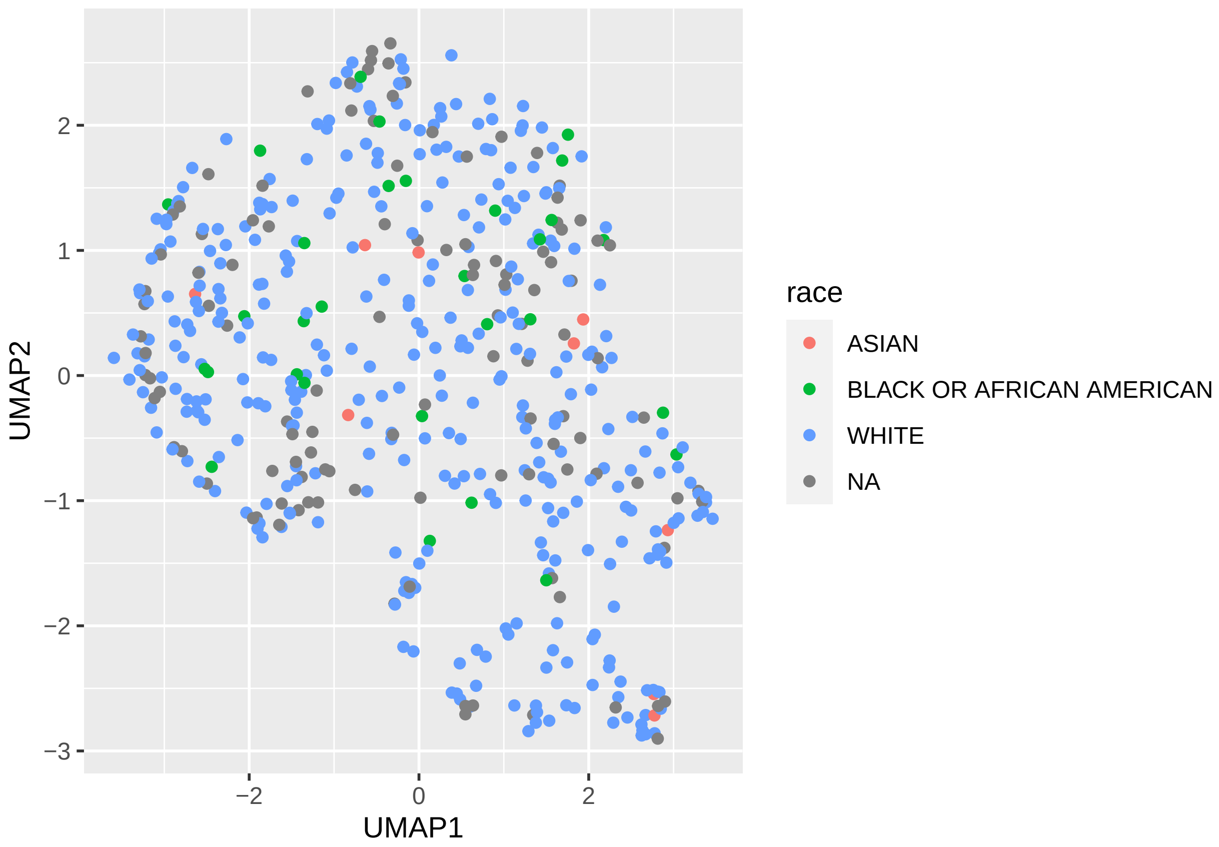
```
lusc_mat <- lusc_fpkm %>%
  column_to_rownames(var = "gene") %>%
  as.matrix() %>%
  t()
```

UMAP analysis across LUSC cohort

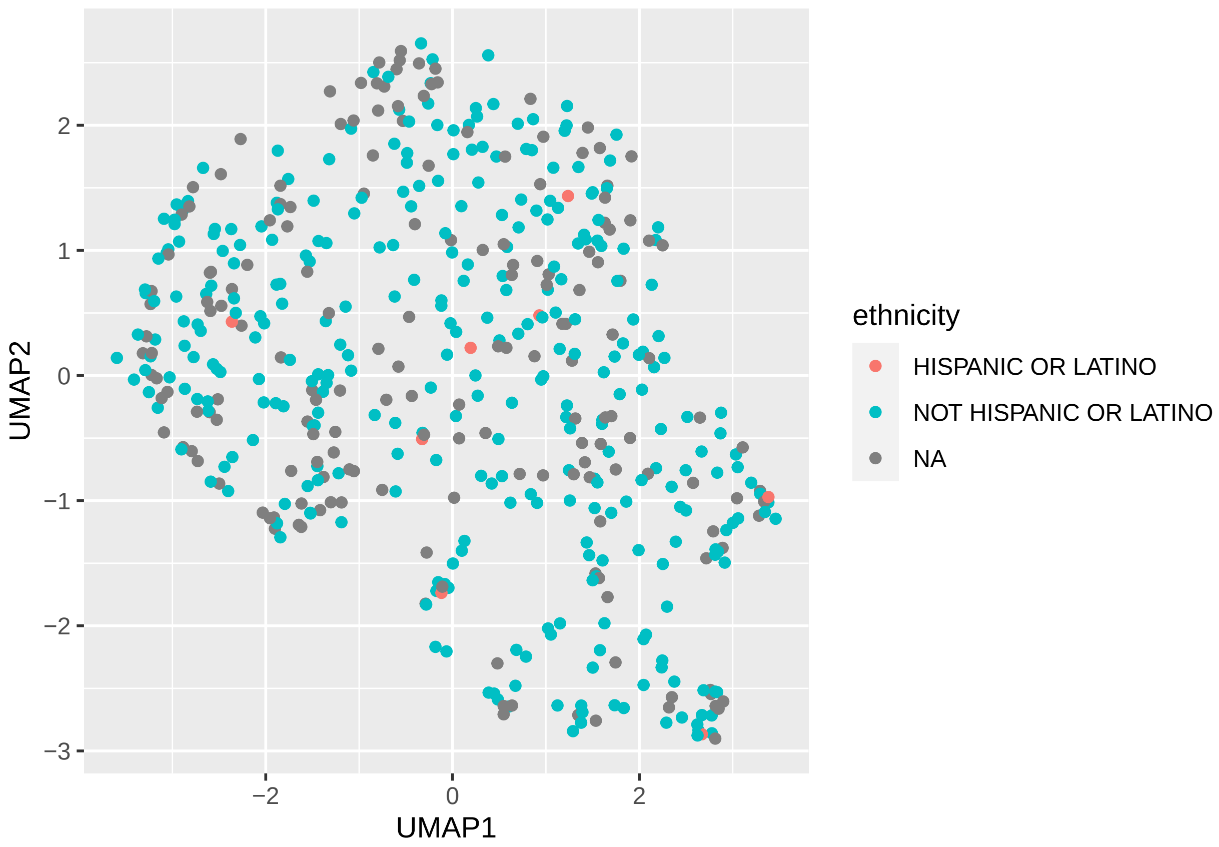
```
lusc_umap <- umap(lusc_mat)

lusc_umap_temp_plot <- lusc_umap %>%
  pluck("layout") %>%
  as.data.frame() %>%
  rownames_to_column(var = "sample") %>%
  as_tibble() %>%
  rename(UMAP1 = V1, UMAP2 = V2) %>%
  inner_join(lusc_metadata, by = "sample") %>%
  ggplot(mapping = aes(x = UMAP1, y = UMAP2))

lusc_umap_plot <- lusc_umap_temp_plot +
  geom_point(mapping = aes(color = gender))
lusc_umap_plot
```



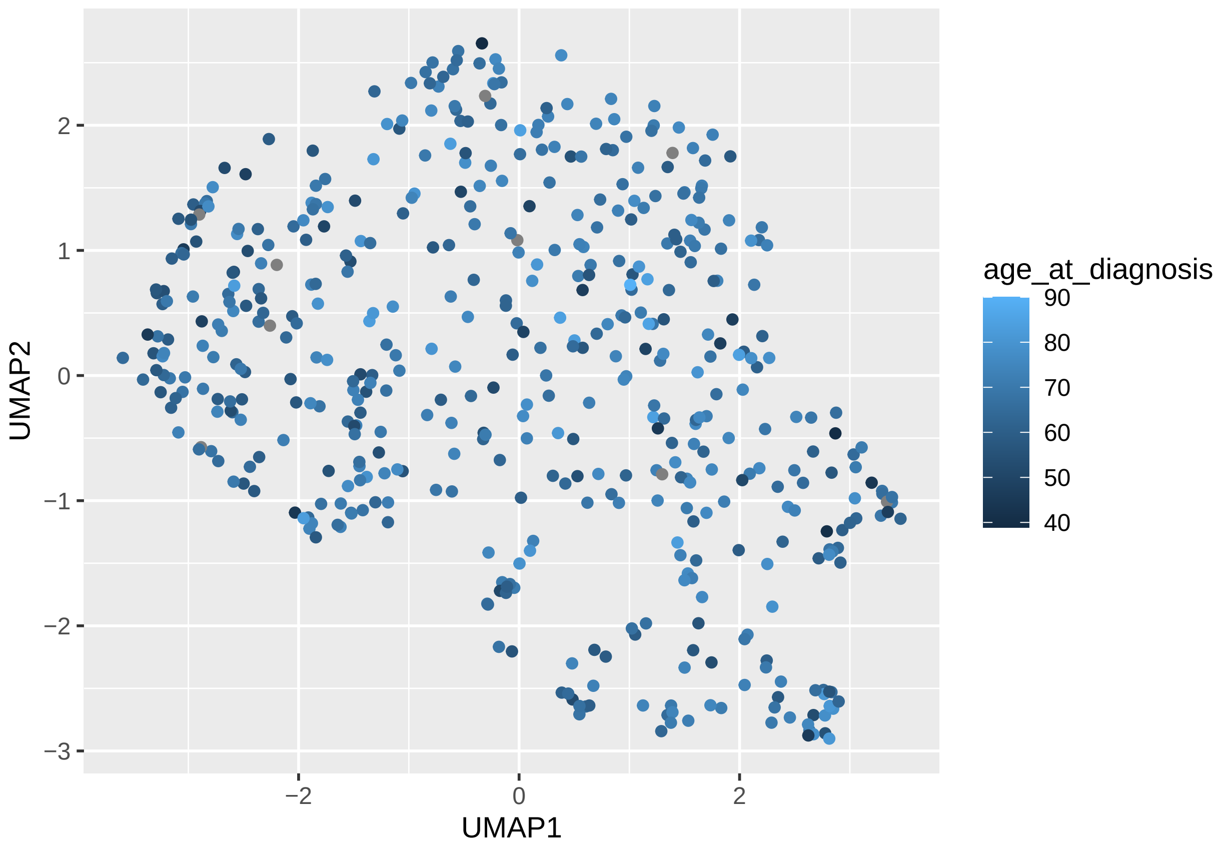
```
lusc_umap_plot <- lusc_umap_temp_plot +
  geom_point(mapping = aes(color = race))
lusc_umap_plot
```



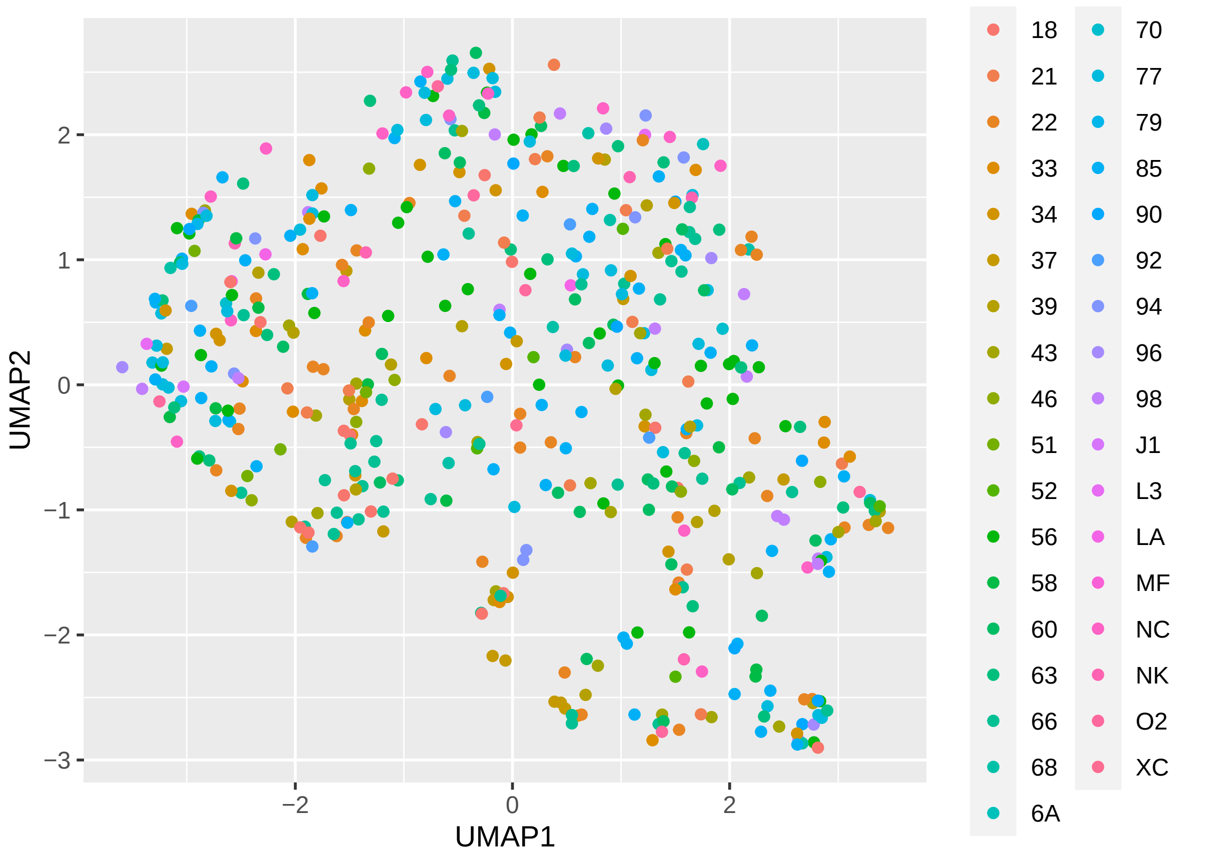
```
lusc_umap_plot <- lusc_umap_temp_plot +
  geom_point(mapping = aes(color = ethnicity))
lusc_umap_plot
```



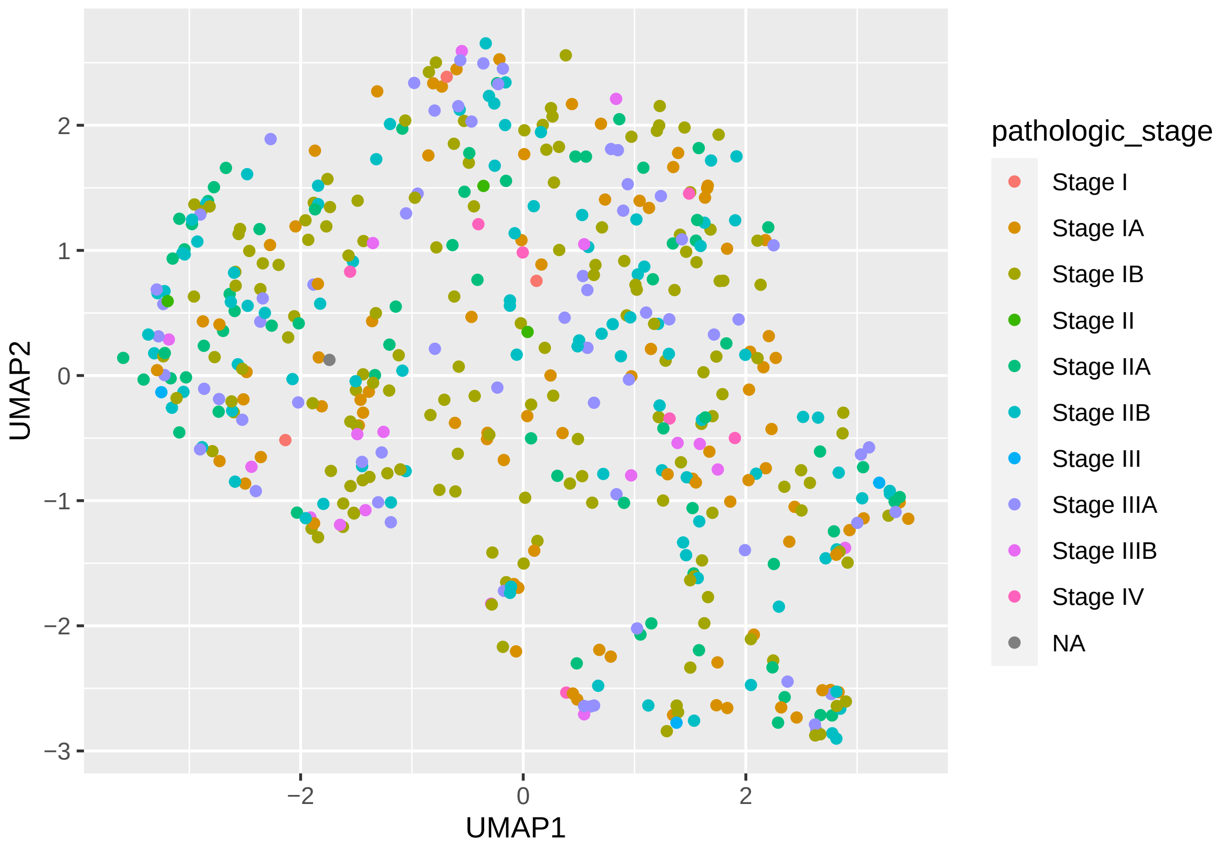
```
lusc_umap_plot <- lusc_umap_temp_plot +
  geom_point(mapping = aes(color = age_at_diagnosis))
lusc_umap_plot
```



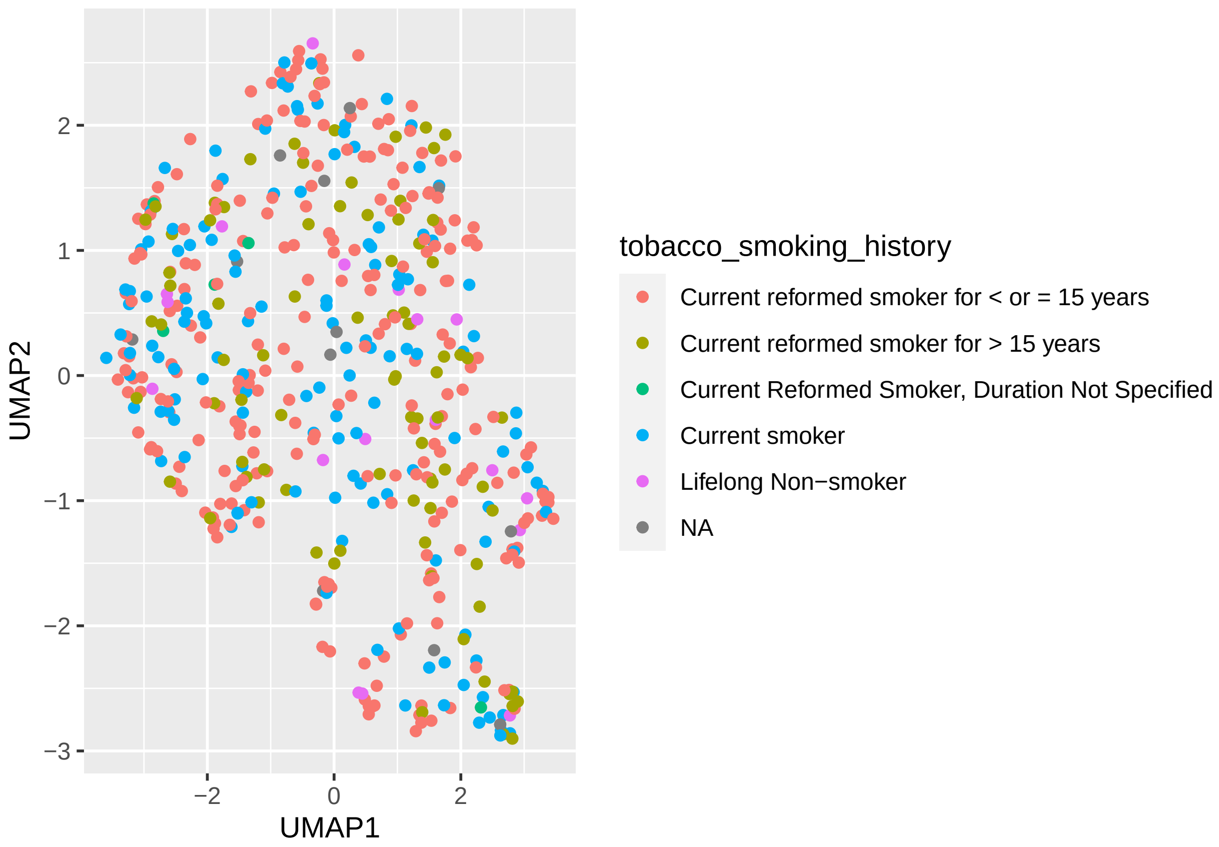
```
lusc_umap_plot <- lusc_umap_temp_plot +
  geom_point(mapping = aes(color = tissue_source_site))
lusc_umap_plot
```



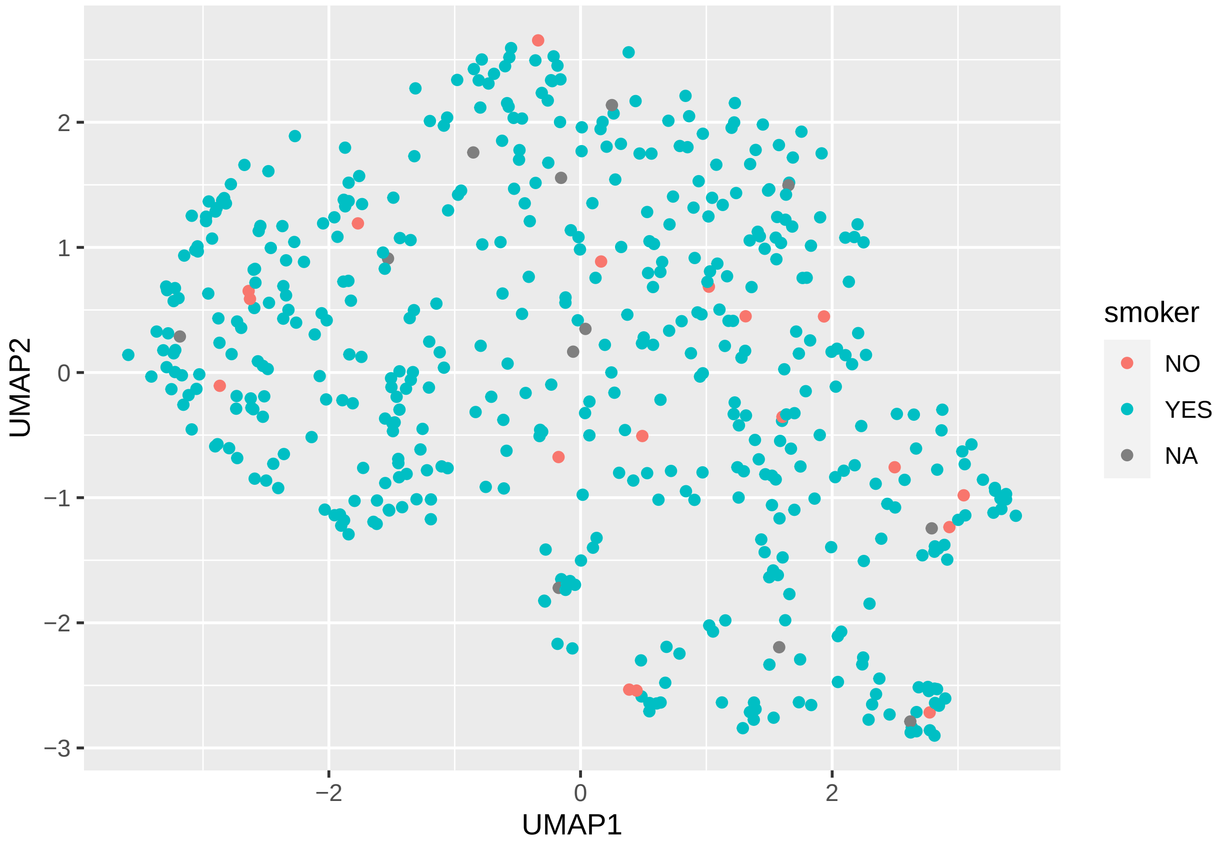
```
lusc_umap_plot <- lusc_umap_temp_plot +
  geom_point(mapping = aes(color = pathologic_stage))
lusc_umap_plot
```



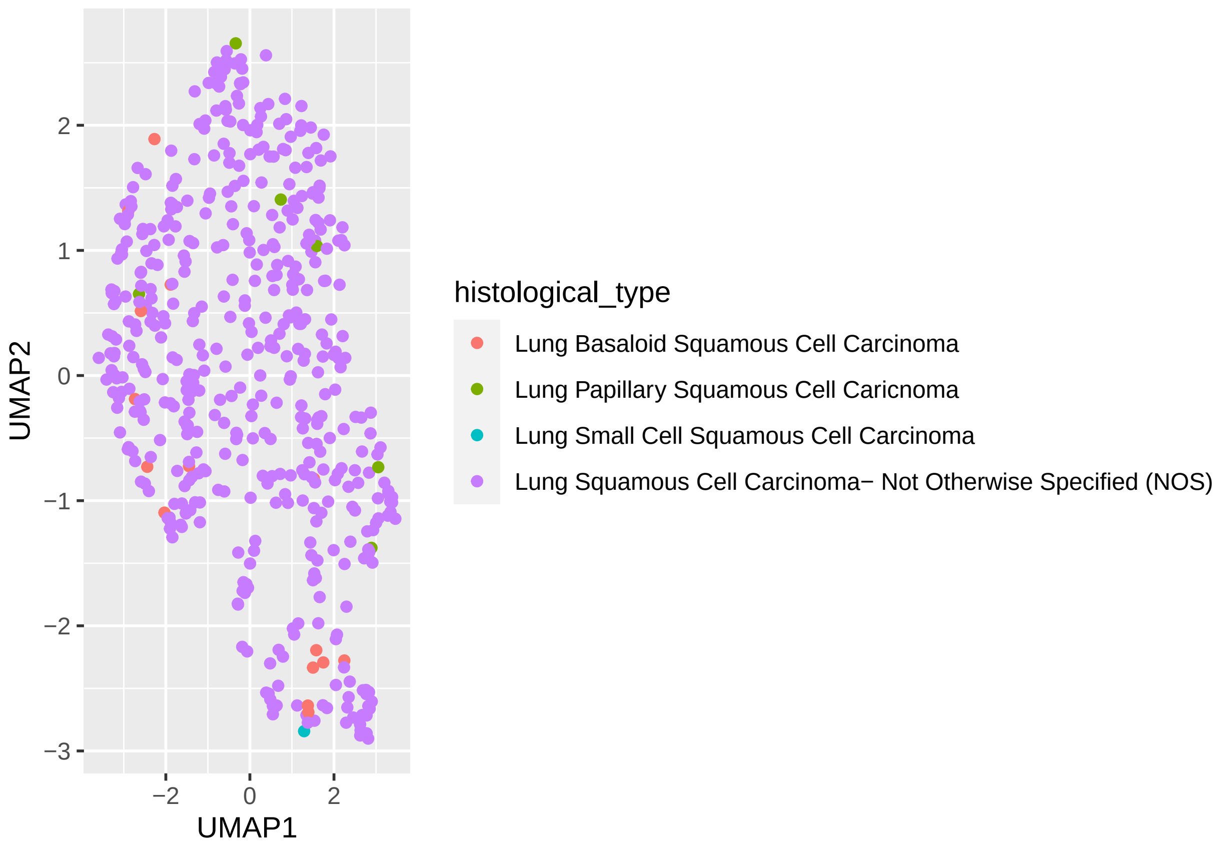
```
lusc_umap_plot <- lusc_umap_temp_plot +
  geom_point(mapping = aes(color = tobacco_smoking_history))
lusc_umap_plot
```



```
lusc_umap_plot <- lusc_umap_temp_plot +
  geom_point(mapping = aes(color = smoker))
lusc_umap_plot
```

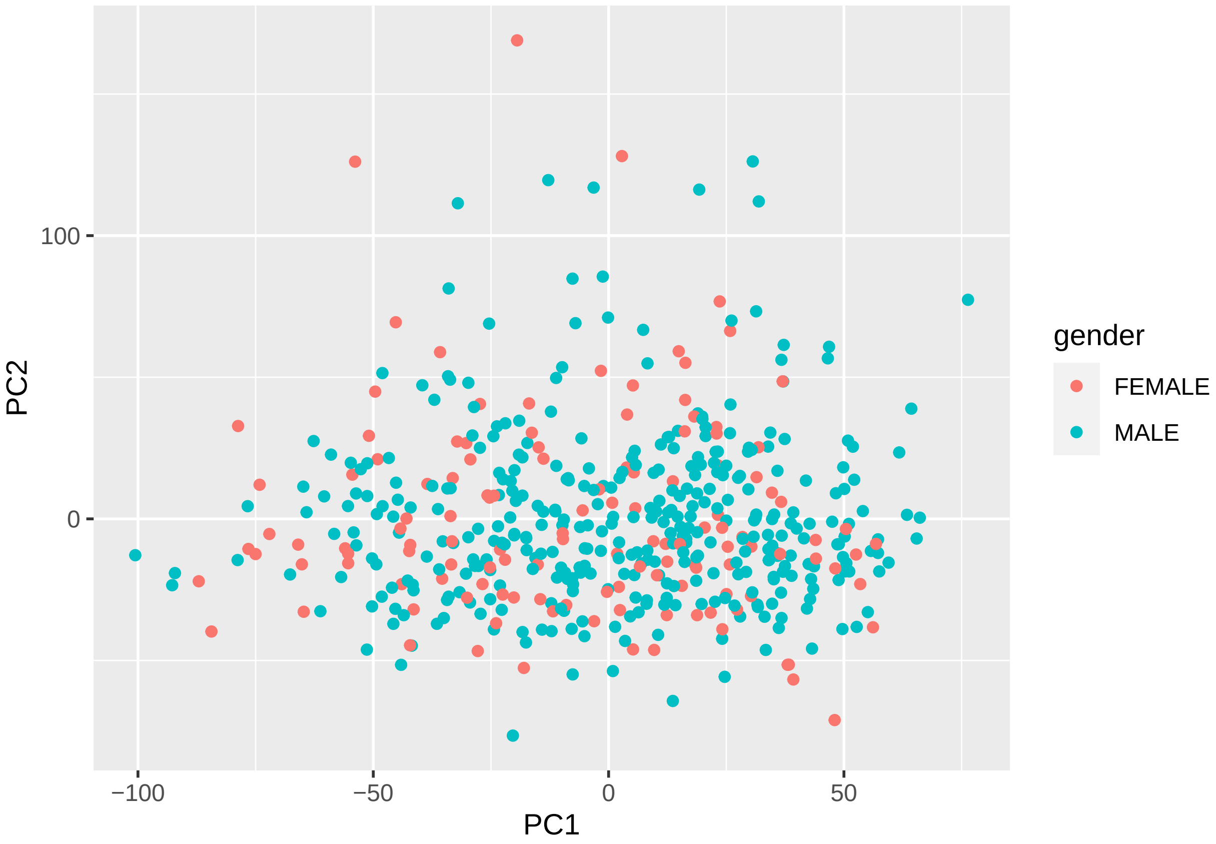


```
lusc_umap_plot <- lusc_umap_temp_plot +
  geom_point(mapping = aes(color = histological_type))
lusc_umap_plot
```

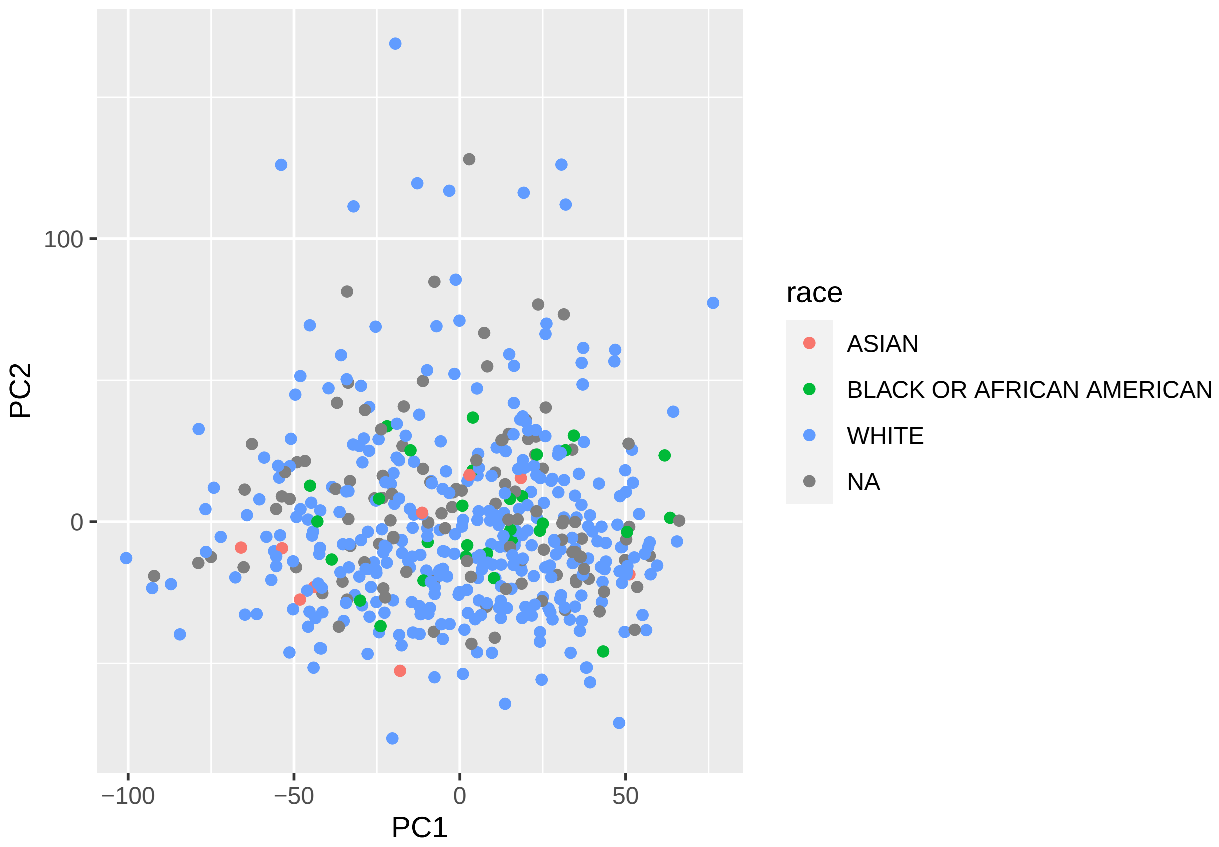


PCA analysis across LUSC cohort

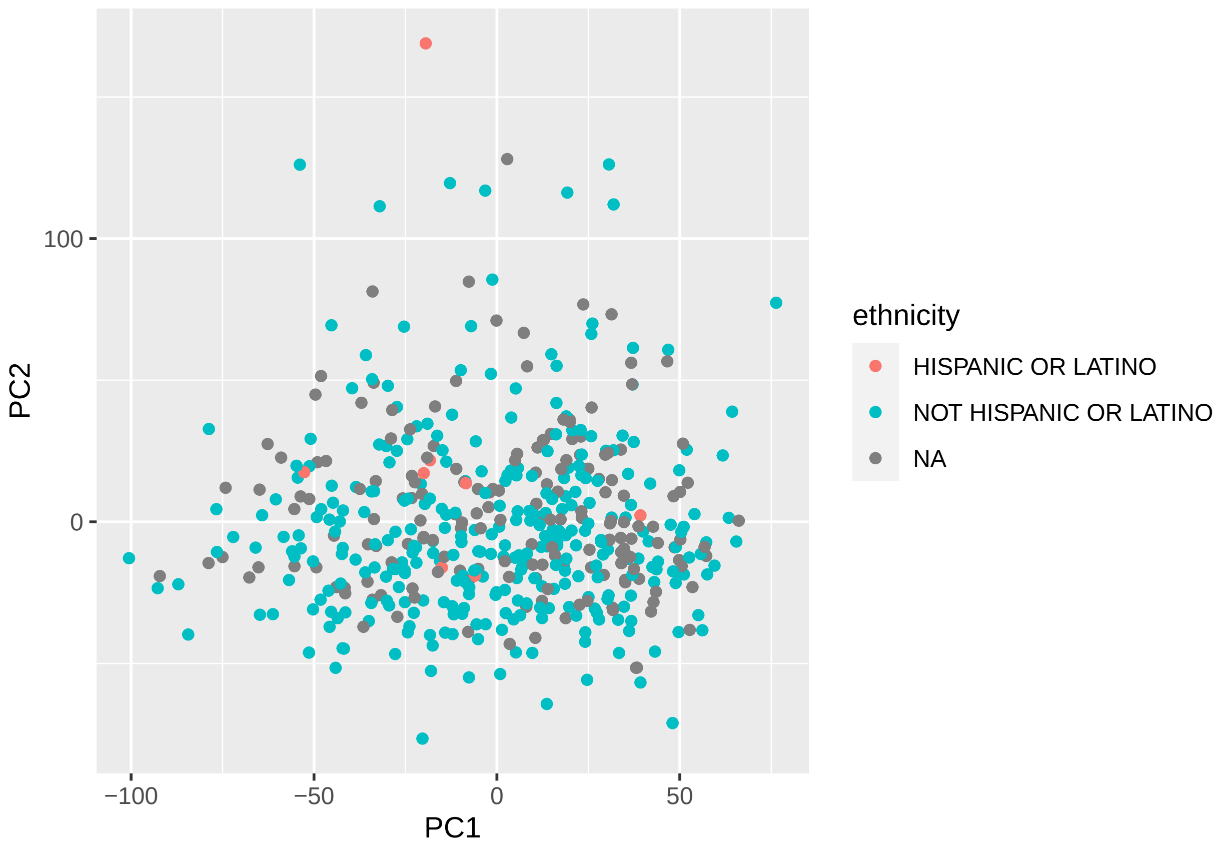
```
lusc_pca <- prcomp(lusc_mat, center = T, scale. = T)

lusc_pca_temp_plot <- lusc_pca %>%
  pluck("x") %>%
  as.data.frame() %>%
  rownames_to_column(var = "sample") %>%
  as_tibble() %>%
  inner_join(lusc_metadata, by = "sample") %>%
  ggplot(mapping = aes(x = PC1, y = PC2))

lusc_pca_plot <- lusc_pca_temp_plot +
  geom_point(mapping = aes(color = gender))
lusc_pca_plot
```



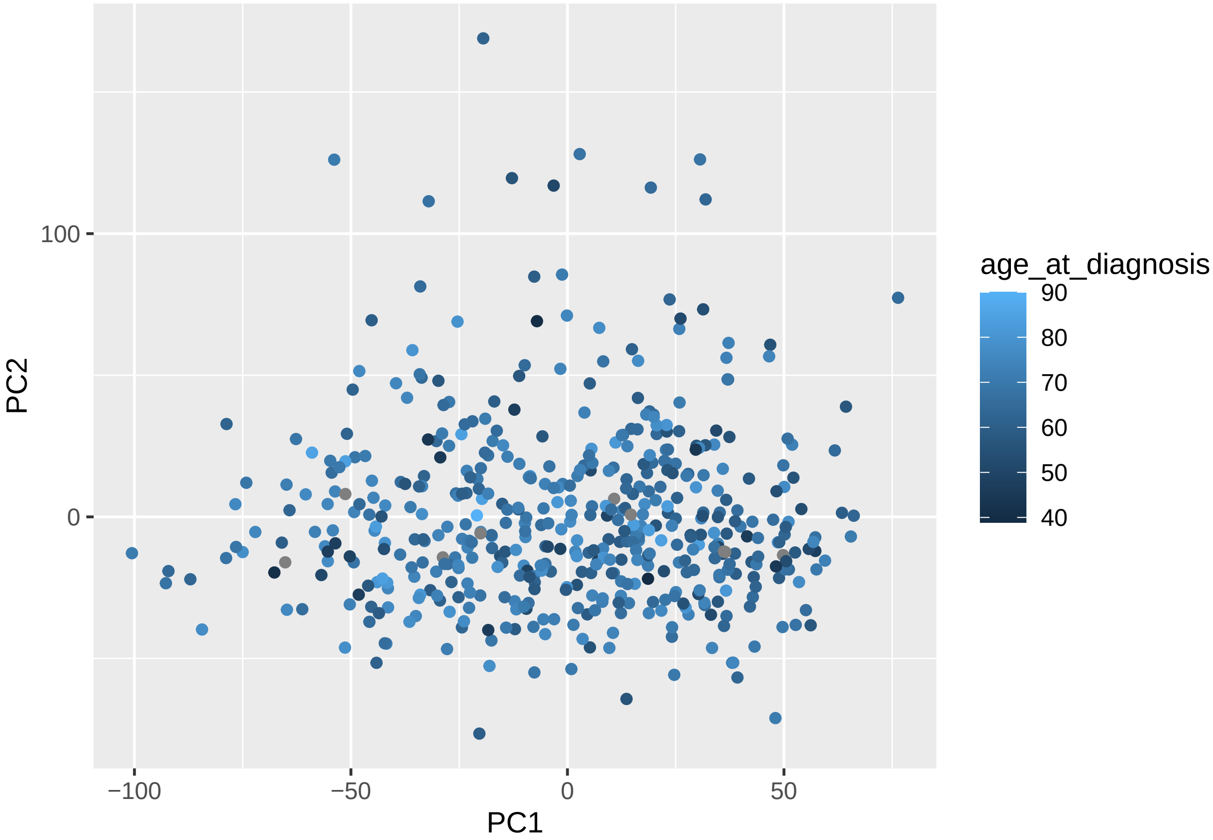
```
lusc_pca_plot <- lusc_pca_temp_plot +
  geom_point(mapping = aes(color = race))
lusc_pca_plot
```



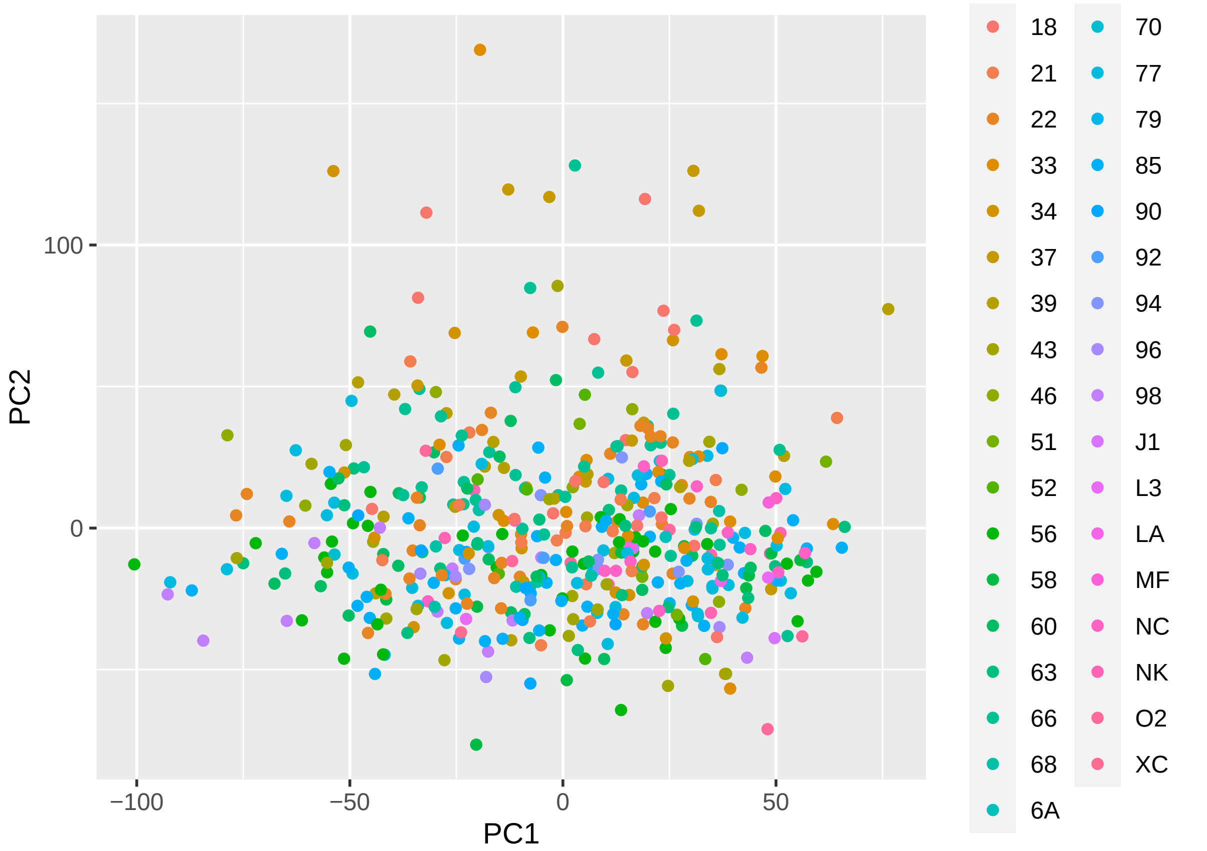
```
lusc_pca_plot <- lusc_pca_temp_plot +
  geom_point(mapping = aes(color = ethnicity))
lusc_pca_plot
```



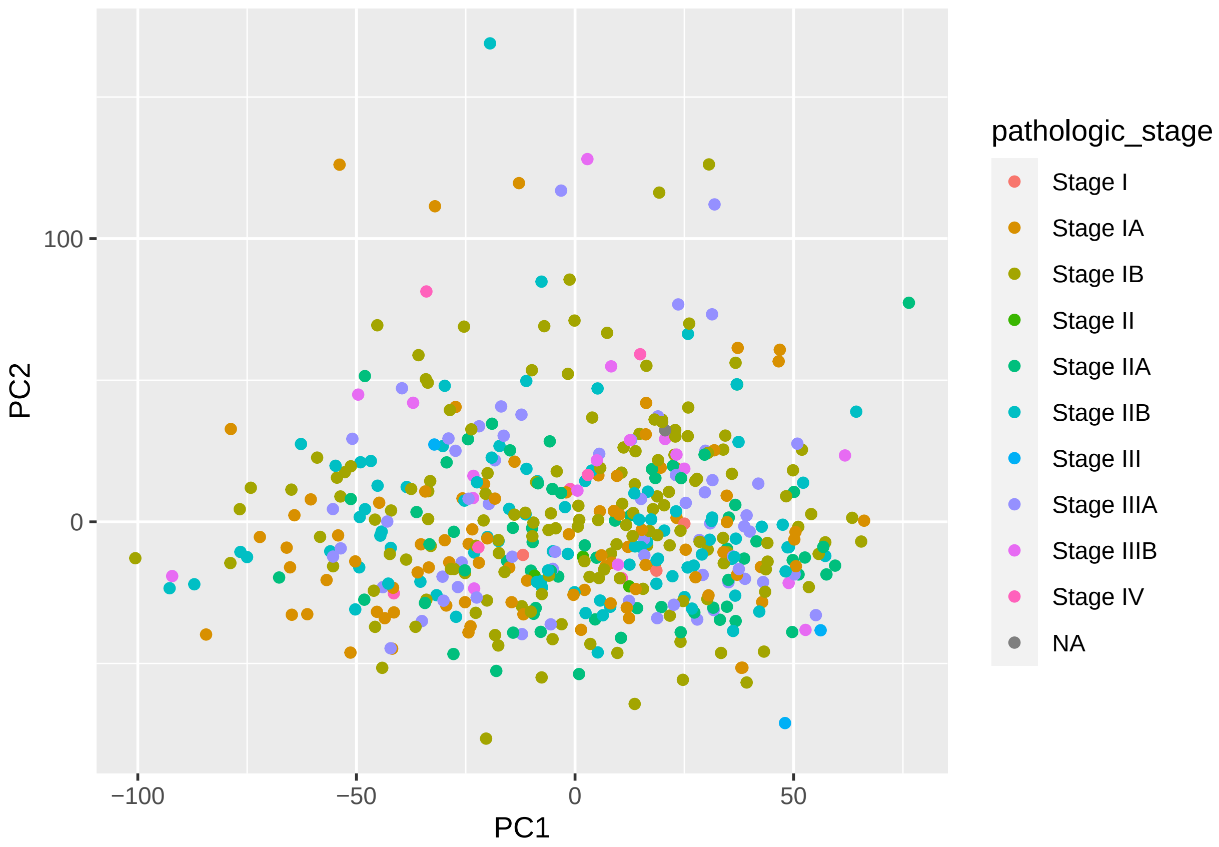
```
lusc_pca_plot <- lusc_pca_temp_plot +
  geom_point(mapping = aes(color = age_at_diagnosis))
lusc_pca_plot
```



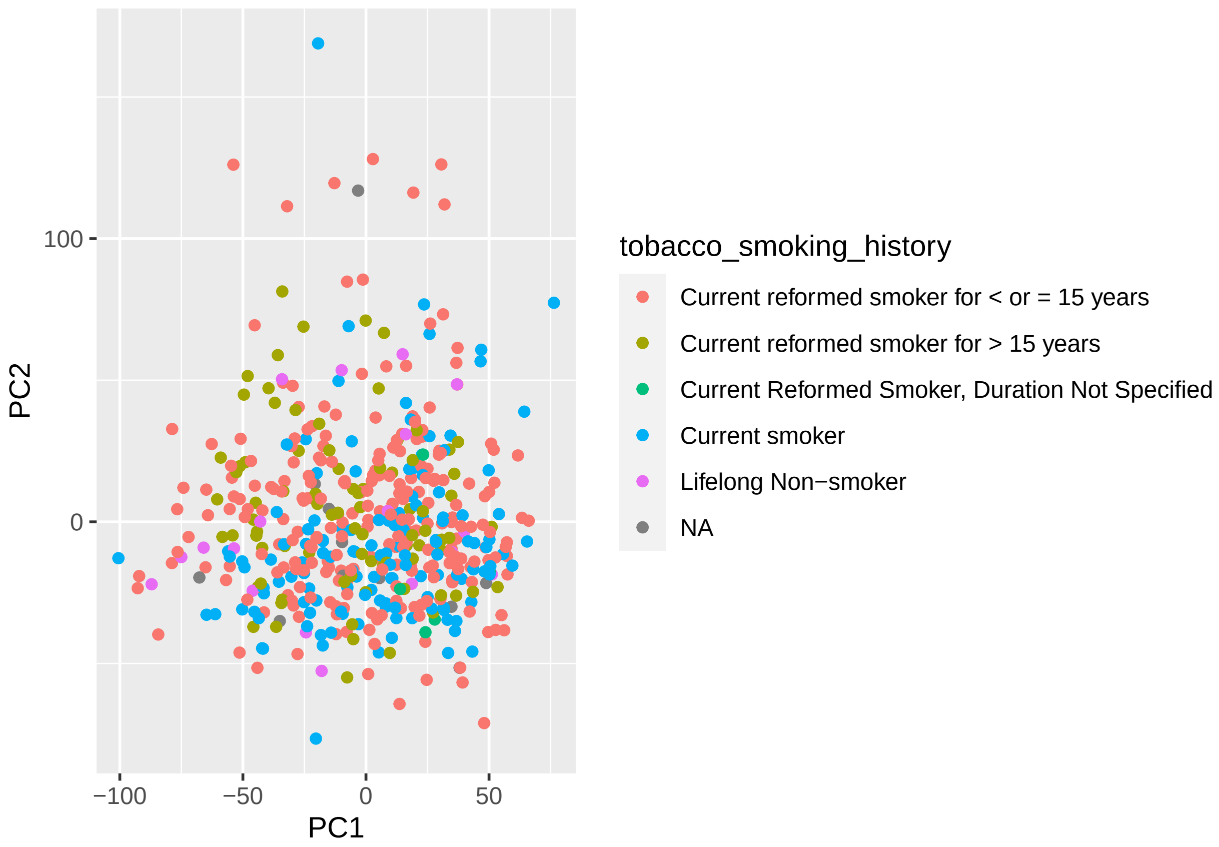
```
lusc_pca_plot <- lusc_pca_temp_plot +
  geom_point(mapping = aes(color = tissue_source_site))
lusc_pca_plot
```



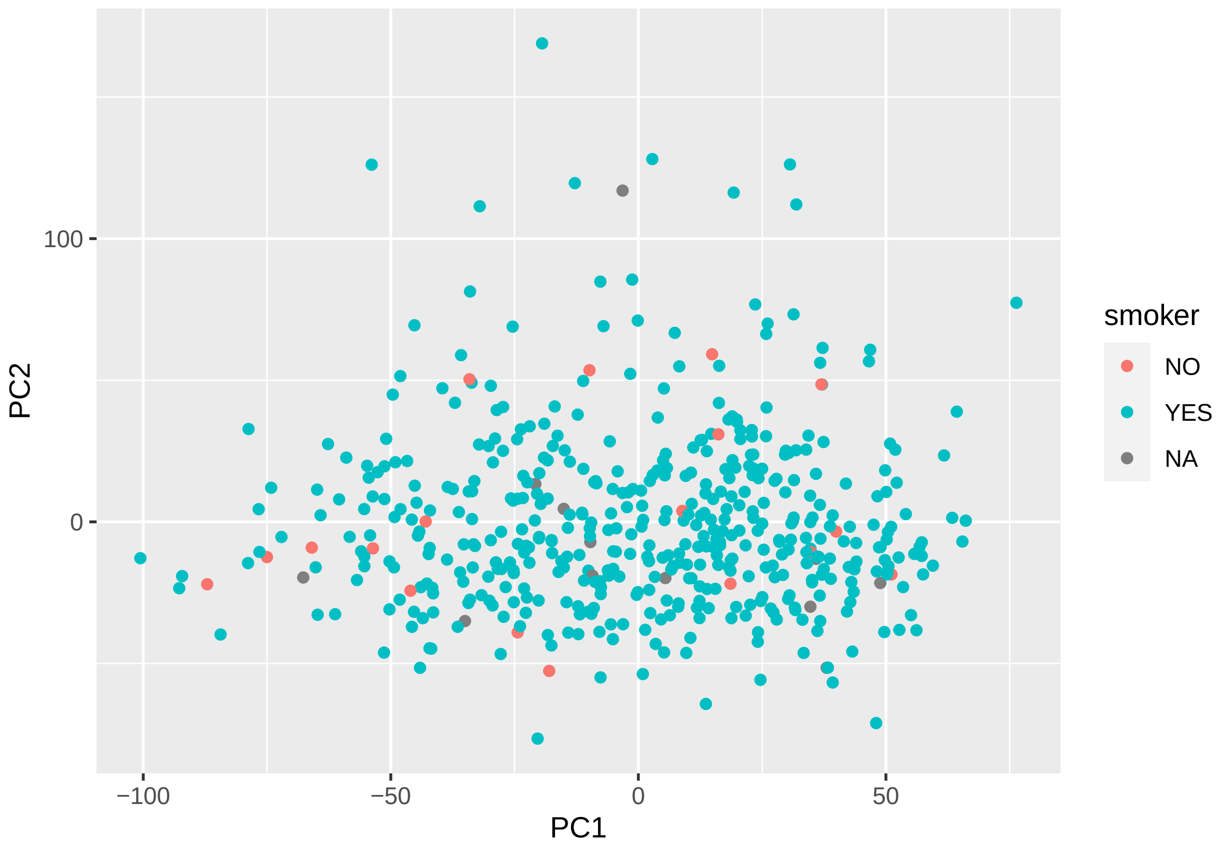
```r
lusc_pca_plot <- lusc_pca_temp_plot +
  geom_point(mapping = aes(color = pathologic_stage))
lusc_pca_plot
```



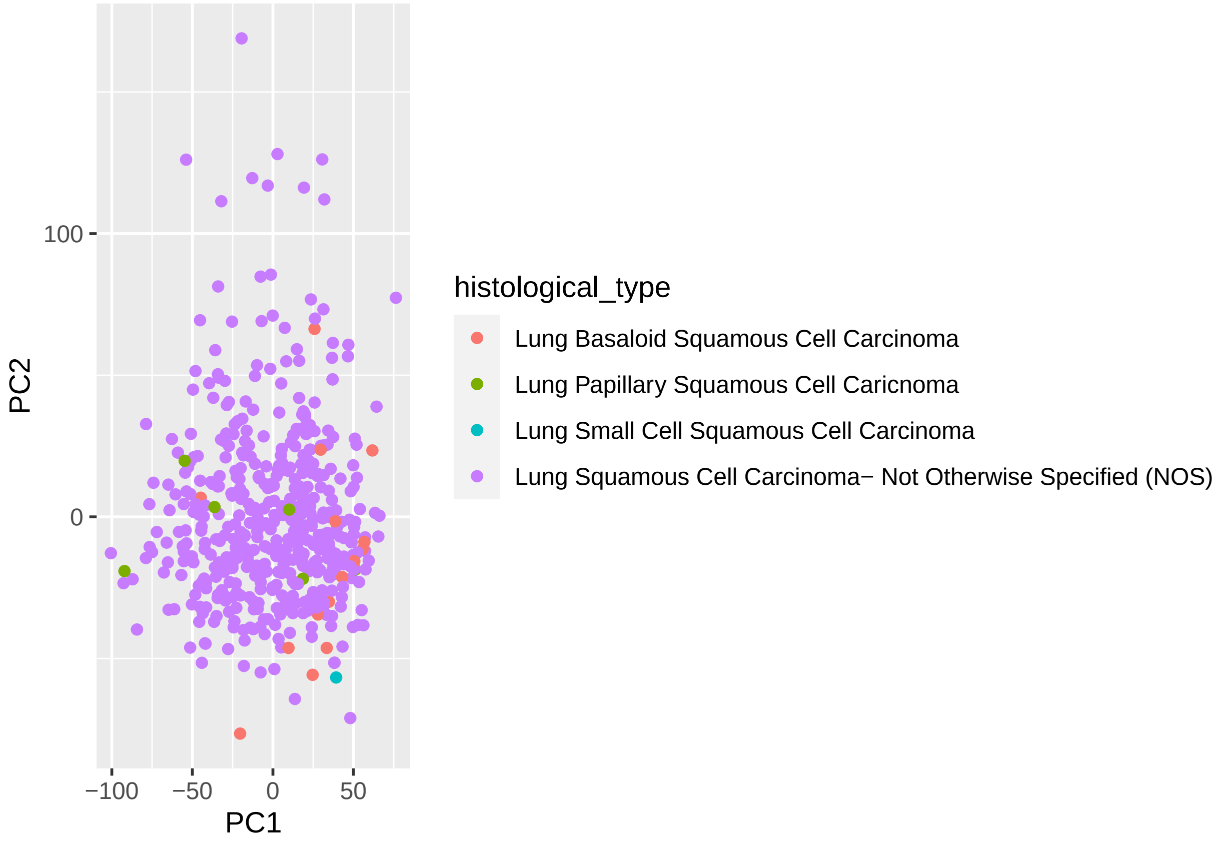
```
lusc_pca_plot <- lusc_pca_temp_plot +
  geom_point(mapping = aes(color = tobacco_smoking_history))
lusc_pca_plot
```



```
lusc_pca_plot <- lusc_pca_temp_plot +
  geom_point(mapping = aes(color = smoker))
lusc_pca_plot
```

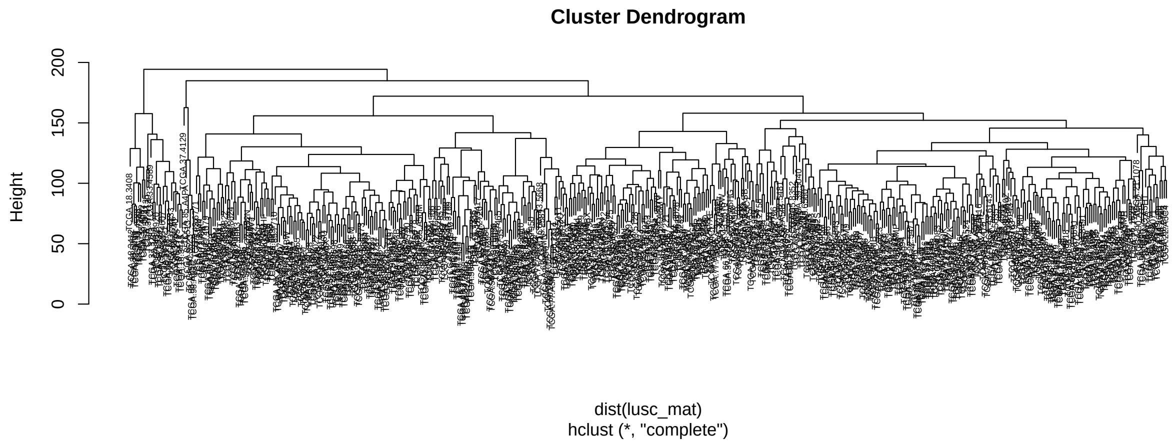


```
lusc_pca_plot <- lusc_pca_temp_plot +
  geom_point(mapping = aes(color = histological_type))
lusc_pca_plot
```

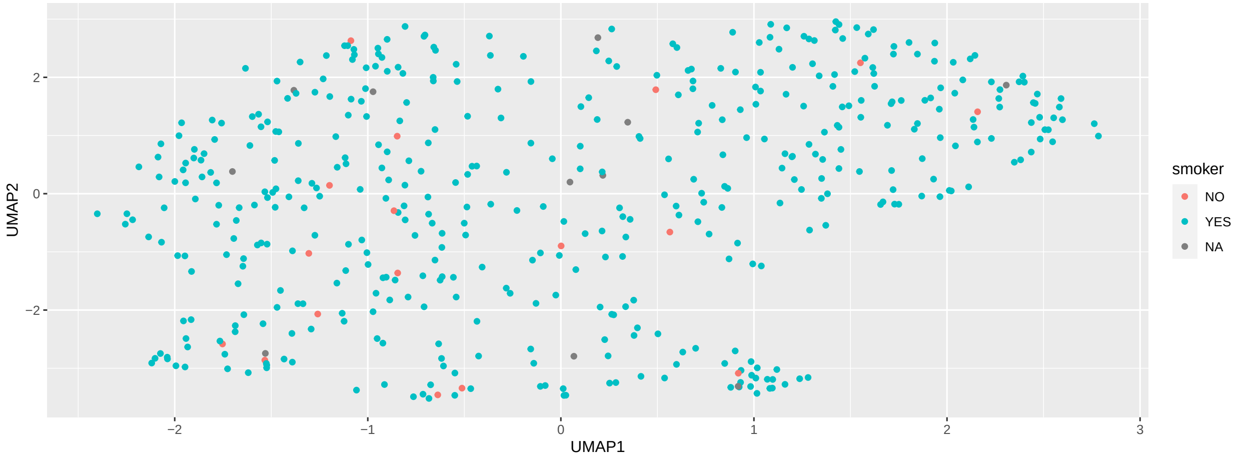


hierachical clustering to select outlier samples use default distance measure (eucledian) and agglomeration method (complete)

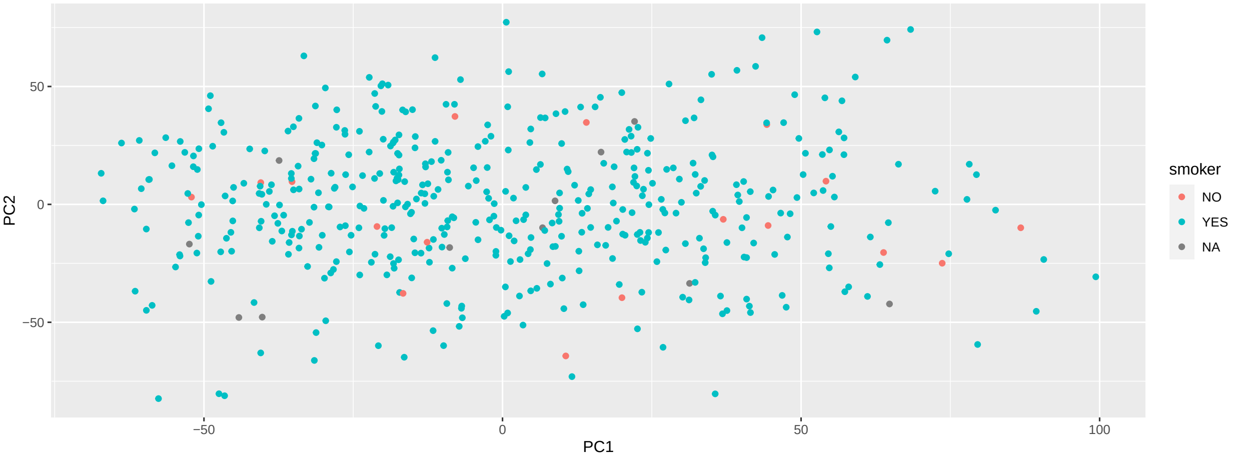
```
lusc_hc <- hclust(dist(lusc_mat))

plot(lusc_hc, cex = 0.5)
```



```
lusc_clusters <- cutree(lusc_hc, h = 190)
lusc_clusters <- tibble(sample = names(lusc_clusters), cluster = unname(lusc_clusters))

write_tsv(x = lusc_clusters, file = "./output/files/02_exploratory_analysis/tcga_lusc_outlier_samples_h
lusc_umap2 <- umap(lusc_mat[rownames(lusc_mat) %in% lusc_clusters[lusc_clusters$cluster == 1, ]$sample,

lusc_umap_temp_plot2 <- lusc_umap2 %>%
  pluck("layout") %>%
  as.data.frame() %>%
  rownames_to_column(var = "sample") %>%
  as_tibble() %>%
  rename(UMAP1 = V1, UMAP2 = V2) %>%
  inner_join(lusc_metadata, by = "sample") %>%
  ggplot(mapping = aes(x = UMAP1, y = UMAP2))

lusc_umap_plot2 <- lusc_umap_temp_plot2 +
  geom_point(mapping = aes(color = smoker))
lusc_umap_plot2
```



```
lusc_pca2 <- prcomp(lusc_mat[rownames(lusc_mat) %in% lusc_clusters[lusc_clusters$cluster == 1, ]$sample

lusc_pca_temp_plot2 <- lusc_pca2 %>%
  pluck("x") %>%
  as.data.frame() %>%
  rownames_to_column(var = "sample") %>%
  as_tibble() %>%
  inner_join(lusc_metadata, by = "sample") %>%
  ggplot(mapping = aes(x = PC1, y = PC2))

lusc_pca_plot2 <- lusc_pca_temp_plot2 +
  geom_point(mapping = aes(color = smoker))
lusc_pca_plot2
```


smoker
NO
YES
NA
PC2
50
0
−50
PC1
−50
0
50
100