NCI-UMD Cohort

9/24/2020

source("code/load\_packages.R")  
source("code/get\_data.R")

# NCI

ncidata <- microbiome.data.nci.umd$meta.dat  
  
# nrow(ncidata)  
# table(ncidata$tissue)  
# table(ncidata$gender)  
# table(ncidata$race)  
# summary(ncidata$age)  
# table(ncidata$bmi)  
# table(ncidata$smoke)  
  
ncidata=ncidata %>%  
 mutate(sample\_type=0,  
 sample\_type=ifelse(tissue=="T" & histology=="ADC",   
 "Esophageal adeno-carcinoma tissues",   
 sample\_type),  
 sample\_type=ifelse(tissue=="T" & histology=="SCC",   
 "Esophageal squamous cell carcinoma tissue",   
 sample\_type),  
 sample\_type=ifelse(tissue=="N" & histology=="ADC",   
 "Esophageal adeno-carcinoma-adjacent tissues",   
 sample\_type),  
 sample\_type=ifelse(tissue=="N" & histology=="SCC",   
 "Esophageal squamous cell carcinoma-adjacent tissue", sample\_type),  
 sample\_type=ifelse(tissue=="BO" & histology=="Barrets only",   
 "Barrett's Esophagus tissue",   
 sample\_type)  
 )  
table(ncidata$sample\_type)

##   
## Barrett's Esophagus tissue   
## 8   
## Esophageal adeno-carcinoma tissues   
## 74   
## Esophageal adeno-carcinoma-adjacent tissues   
## 87   
## Esophageal squamous cell carcinoma tissue   
## 17   
## Esophageal squamous cell carcinoma-adjacent tissue   
## 27

# ncidata %>%  
# group\_by(diagnosis) %>%  
# summarize(mean\_shannon = mean(shannon), sd\_shannon = sd(shannon))

N=table(ncidata$sample\_type)  
  
sex1=ncidata %>%  
 group\_by(sample\_type, gender)%>%  
 summarize(N=n())%>%  
 ungroup(gender)%>%  
 mutate(p=N/sum(N))%>%  
 ungroup(sample\_type)%>%  
 group\_by(gender)%>%  
 mutate(t=sum(N))%>%  
 ungroup(gender)%>%  
 mutate(pt=t/sum(N))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

sex1

## # A tibble: 9 x 6  
## sample\_type gender N p t pt  
## <chr> <chr> <int> <dbl> <int> <dbl>  
## 1 Barrett's Esophagus tissue M 8 1 173 0.812  
## 2 Esophageal adeno-carcinoma tissues F 9 0.122 40 0.188  
## 3 Esophageal adeno-carcinoma tissues M 65 0.878 173 0.812  
## 4 Esophageal adeno-carcinoma-adjacent tissues F 10 0.115 40 0.188  
## 5 Esophageal adeno-carcinoma-adjacent tissues M 77 0.885 173 0.812  
## 6 Esophageal squamous cell carcinoma tissue F 8 0.471 40 0.188  
## 7 Esophageal squamous cell carcinoma tissue M 9 0.529 173 0.812  
## 8 Esophageal squamous cell carcinoma-adjacent ti… F 13 0.481 40 0.188  
## 9 Esophageal squamous cell carcinoma-adjacent ti… M 14 0.519 173 0.812

race1=ncidata %>%  
 group\_by(sample\_type, race)%>%  
 summarize(N=n())%>%  
 ungroup(race)%>%  
 mutate(p=N/sum(N))%>%  
 ungroup(sample\_type)%>%  
 group\_by(race)%>%  
 mutate(t=sum(N))%>%  
 ungroup(race)%>%  
 mutate(pt=t/sum(N))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

race1

## # A tibble: 13 x 6  
## sample\_type race N p t pt  
## <chr> <chr> <int> <dbl> <int> <dbl>  
## 1 Barrett's Esophagus tissue W 8 1 191 0.897   
## 2 Esophageal adeno-carcinoma tissues B 1 0.0135 17 0.0798   
## 3 Esophageal adeno-carcinoma tissues H 1 0.0135 2 0.00939  
## 4 Esophageal adeno-carcinoma tissues W 72 0.973 191 0.897   
## 5 Esophageal adeno-carcinoma-adjacent tissues B 1 0.0115 17 0.0798   
## 6 Esophageal adeno-carcinoma-adjacent tissues H 1 0.0115 2 0.00939  
## 7 Esophageal adeno-carcinoma-adjacent tissues W 85 0.977 191 0.897   
## 8 Esophageal squamous cell carcinoma tissue B 6 0.353 17 0.0798   
## 9 Esophageal squamous cell carcinoma tissue O 1 0.0588 3 0.0141   
## 10 Esophageal squamous cell carcinoma tissue W 10 0.588 191 0.897   
## 11 Esophageal squamous cell carcinoma-adjacent… B 9 0.333 17 0.0798   
## 12 Esophageal squamous cell carcinoma-adjacent… O 2 0.0741 3 0.0141   
## 13 Esophageal squamous cell carcinoma-adjacent… W 16 0.593 191 0.897

age1=ncidata %>%  
 group\_by(sample\_type)%>%  
 summarize(mean=mean(age),  
 sd=sd(age))

## `summarise()` ungrouping output (override with `.groups` argument)

age1

## # A tibble: 5 x 3  
## sample\_type mean sd  
## <chr> <dbl> <dbl>  
## 1 Barrett's Esophagus tissue 65.9 6.03  
## 2 Esophageal adeno-carcinoma tissues 60.3 10.6   
## 3 Esophageal adeno-carcinoma-adjacent tissues 61.9 11.0   
## 4 Esophageal squamous cell carcinoma tissue 60.7 9.31  
## 5 Esophageal squamous cell carcinoma-adjacent tissue 58.7 9.19

mean(ncidata$age)

## [1] 60.98122

sd(ncidata$age)

## [1] 10.41632

bmi1=ncidata %>%  
 group\_by(sample\_type, bmicat)%>%  
 summarize(N=n())%>%  
 ungroup(bmicat)%>%  
 mutate(p=N/sum(N))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

bmi1

## # A tibble: 23 x 4  
## # Groups: sample\_type [5]  
## sample\_type bmicat N p  
## <chr> <chr> <int> <dbl>  
## 1 Barrett's Esophagus tissue normal weight 1 0.125   
## 2 Barrett's Esophagus tissue obese 3 0.375   
## 3 Barrett's Esophagus tissue overweight 2 0.25   
## 4 Barrett's Esophagus tissue U 2 0.25   
## 5 Esophageal adeno-carcinoma tissues normal weight 21 0.284   
## 6 Esophageal adeno-carcinoma tissues obese 23 0.311   
## 7 Esophageal adeno-carcinoma tissues overweight 24 0.324   
## 8 Esophageal adeno-carcinoma tissues U 5 0.0676  
## 9 Esophageal adeno-carcinoma tissues underweight 1 0.0135  
## 10 Esophageal adeno-carcinoma-adjacent tissues normal weight 19 0.218   
## # … with 13 more rows

barretts1=ncidata %>%  
 group\_by(sample\_type, barretts)%>%  
 summarize(N=n())%>%  
 ungroup(barretts)%>%  
 mutate(p=N/sum(N))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

barretts1

## # A tibble: 9 x 4  
## # Groups: sample\_type [5]  
## sample\_type barretts N p  
## <chr> <chr> <int> <dbl>  
## 1 Barrett's Esophagus tissue Y 8 1   
## 2 Esophageal adeno-carcinoma tissues N 24 0.324  
## 3 Esophageal adeno-carcinoma tissues Y 50 0.676  
## 4 Esophageal adeno-carcinoma-adjacent tissues N 27 0.310  
## 5 Esophageal adeno-carcinoma-adjacent tissues Y 60 0.690  
## 6 Esophageal squamous cell carcinoma tissue N 14 0.824  
## 7 Esophageal squamous cell carcinoma tissue Y 3 0.176  
## 8 Esophageal squamous cell carcinoma-adjacent tissue N 23 0.852  
## 9 Esophageal squamous cell carcinoma-adjacent tissue Y 4 0.148

smoke1=ncidata %>%  
 group\_by(sample\_type, smoke)%>%  
 summarize(N=n())%>%  
 ungroup(smoke)%>%  
 mutate(p=N/sum(N))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

smoke1

## # A tibble: 15 x 4  
## # Groups: sample\_type [5]  
## sample\_type smoke N p  
## <chr> <chr> <int> <dbl>  
## 1 Barrett's Esophagus tissue N 2 0.25   
## 2 Barrett's Esophagus tissue U 1 0.125   
## 3 Barrett's Esophagus tissue Y 5 0.625   
## 4 Esophageal adeno-carcinoma tissues N 15 0.203   
## 5 Esophageal adeno-carcinoma tissues U 3 0.0405  
## 6 Esophageal adeno-carcinoma tissues Y 56 0.757   
## 7 Esophageal adeno-carcinoma-adjacent tissues N 14 0.161   
## 8 Esophageal adeno-carcinoma-adjacent tissues U 9 0.103   
## 9 Esophageal adeno-carcinoma-adjacent tissues Y 64 0.736   
## 10 Esophageal squamous cell carcinoma tissue N 2 0.118   
## 11 Esophageal squamous cell carcinoma tissue U 1 0.0588  
## 12 Esophageal squamous cell carcinoma tissue Y 14 0.824   
## 13 Esophageal squamous cell carcinoma-adjacent tissue N 2 0.0741  
## 14 Esophageal squamous cell carcinoma-adjacent tissue U 3 0.111   
## 15 Esophageal squamous cell carcinoma-adjacent tissue Y 22 0.815

stage1=ncidata %>%  
 group\_by(sample\_type, stage)%>%  
 summarize(N=n())%>%  
 ungroup(stage)%>%  
 mutate(p=N/sum(N))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

stage1

## # A tibble: 24 x 4  
## # Groups: sample\_type [5]  
## sample\_type stage N p  
## <chr> <chr> <int> <dbl>  
## 1 Barrett's Esophagus tissue <NA> 8 1   
## 2 Esophageal adeno-carcinoma tissues 0 2 0.0270  
## 3 Esophageal adeno-carcinoma tissues I 8 0.108   
## 4 Esophageal adeno-carcinoma tissues IIA 27 0.365   
## 5 Esophageal adeno-carcinoma tissues IIB 5 0.0676  
## 6 Esophageal adeno-carcinoma tissues III 26 0.351   
## 7 Esophageal adeno-carcinoma tissues IV 5 0.0676  
## 8 Esophageal adeno-carcinoma tissues <NA> 1 0.0135  
## 9 Esophageal adeno-carcinoma-adjacent tissues 0 3 0.0345  
## 10 Esophageal adeno-carcinoma-adjacent tissues I 12 0.138   
## # … with 14 more rows

neoadj=ncidata %>%  
 group\_by(sample\_type, neoadj\_treatment)%>%  
 summarize(N=n())%>%  
 ungroup(neoadj\_treatment)%>%  
 mutate(p=N/sum(N)) %>%  
 filter(neoadj\_treatment=="Y")

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

neoadj

## # A tibble: 5 x 4  
## # Groups: sample\_type [5]  
## sample\_type neoadj\_treatment N p  
## <chr> <chr> <int> <dbl>  
## 1 Barrett's Esophagus tissue Y 8 1   
## 2 Esophageal adeno-carcinoma tissues Y 44 0.595  
## 3 Esophageal adeno-carcinoma-adjacent tissues Y 52 0.598  
## 4 Esophageal squamous cell carcinoma tissue Y 9 0.529  
## 5 Esophageal squamous cell carcinoma-adjacent tiss… Y 18 0.667

survival=ncidata %>%  
 group\_by(sample\_type)%>%  
 summarize(mean=mean(days\_from\_surg\_to\_event),  
 sd=sd(days\_from\_surg\_to\_event))

## `summarise()` ungrouping output (override with `.groups` argument)

survival

## # A tibble: 5 x 3  
## sample\_type mean sd  
## <chr> <dbl> <dbl>  
## 1 Barrett's Esophagus tissue NA NA   
## 2 Esophageal adeno-carcinoma tissues 1923. 1813.  
## 3 Esophageal adeno-carcinoma-adjacent tissues 1849. 1754.  
## 4 Esophageal squamous cell carcinoma tissue 1924 1872.  
## 5 Esophageal squamous cell carcinoma-adjacent tissue 1699. 1911.

# TCGA RNA seq

tcga\_rna\_seq <- microbiome.data.tcga.RNAseq$meta.dat  
  
tcga\_rna\_seq=tcga\_rna\_seq %>%  
 mutate(sample\_type=0,  
 sample\_type=ifelse((morphology=="8140/3" |  
 morphology=="8480/3") &  
 SampleType\_Level2=="Tumor",  
 "Esophageal adeno-carcinoma tissues",  
 sample\_type),  
 sample\_type=ifelse((morphology=="8140/3" |   
 morphology=="8480/3") &  
 SampleType\_Level2=="Normal",  
 "Esophageal adeno-carcinoma-adjacent tissues",   
 sample\_type),  
 sample\_type=ifelse((morphology=="8070/3" |   
 morphology=="8071/3" |   
 morphology=="8083/3") &  
 SampleType\_Level2=="Tumor",  
 "Esophageal squamous cell carcinoma tissue",   
 sample\_type),  
 sample\_type=ifelse((morphology=="8070/3" |  
 morphology=="8071/3" |  
 morphology=="8083/3") &  
 SampleType\_Level2=="Normal",  
 "Esophageal squamous cell carcinoma-adjacent tissue",   
 sample\_type),  
 sample\_type=ifelse((morphology=="8070/3" |  
 morphology=="8071/3" |  
 morphology=="8083/3") &  
 SampleType\_Level2=="Normal",  
 "Esophageal squamous cell carcinoma-adjacent tissue",   
 sample\_type)  
 )  
table(tcga\_rna\_seq$sample\_type)

##   
## Esophageal adeno-carcinoma tissues   
## 28   
## Esophageal adeno-carcinoma-adjacent tissues   
## 7   
## Esophageal squamous cell carcinoma tissue   
## 31

N2=table(tcga\_rna\_seq$sample\_type)  
N2

##   
## Esophageal adeno-carcinoma tissues   
## 28   
## Esophageal adeno-carcinoma-adjacent tissues   
## 7   
## Esophageal squamous cell carcinoma tissue   
## 31

sex2=tcga\_rna\_seq %>%  
 group\_by(sample\_type, Gender)%>%  
 summarize(N2=n())%>%  
 ungroup(Gender)%>%  
 mutate(p=N2/sum(N2))%>%  
 ungroup(sample\_type)%>%  
 group\_by(Gender)%>%  
 mutate(t=sum(N2))%>%  
 ungroup(Gender)%>%  
 mutate(pt=t/sum(N2))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

sex2

## # A tibble: 6 x 6  
## sample\_type Gender N2 p t pt  
## <chr> <chr> <int> <dbl> <int> <dbl>  
## 1 Esophageal adeno-carcinoma tissues female 6 0.214 11 0.167  
## 2 Esophageal adeno-carcinoma tissues male 22 0.786 55 0.833  
## 3 Esophageal adeno-carcinoma-adjacent tissues female 2 0.286 11 0.167  
## 4 Esophageal adeno-carcinoma-adjacent tissues male 5 0.714 55 0.833  
## 5 Esophageal squamous cell carcinoma tissue female 3 0.0968 11 0.167  
## 6 Esophageal squamous cell carcinoma tissue male 28 0.903 55 0.833

race2=tcga\_rna\_seq %>%  
 group\_by(sample\_type, race)%>%  
 summarize(N2=n())%>%  
 ungroup(race)%>%  
 mutate(p=N2/sum(N2))%>%  
 ungroup(sample\_type)%>%  
 group\_by(race)%>%  
 mutate(t=sum(N2))%>%  
 ungroup(race)%>%  
 mutate(pt=t/sum(N2))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

race2

## # A tibble: 7 x 6  
## sample\_type race N2 p t pt  
## <chr> <chr> <int> <dbl> <int> <dbl>  
## 1 Esophageal adeno-carcinoma tissu… asian 1 0.0357 21 0.318   
## 2 Esophageal adeno-carcinoma tissu… not reported 4 0.143 4 0.0606  
## 3 Esophageal adeno-carcinoma tissu… white 23 0.821 40 0.606   
## 4 Esophageal adeno-carcinoma-adjac… white 7 1 40 0.606   
## 5 Esophageal squamous cell carcino… asian 20 0.645 21 0.318   
## 6 Esophageal squamous cell carcino… black or african … 1 0.0323 1 0.0152  
## 7 Esophageal squamous cell carcino… white 10 0.323 40 0.606

age2=tcga\_rna\_seq %>%  
 group\_by(sample\_type)%>%  
 summarize(mean=mean(days\_to\_birth/(-365)),  
 sd=sd(days\_to\_birth/(-365)))

## `summarise()` ungrouping output (override with `.groups` argument)

age2

## # A tibble: 3 x 3  
## sample\_type mean sd  
## <chr> <dbl> <dbl>  
## 1 Esophageal adeno-carcinoma tissues 67.2 13.2   
## 2 Esophageal adeno-carcinoma-adjacent tissues 75.7 6.90  
## 3 Esophageal squamous cell carcinoma tissue 58.6 11.1

mean(tcga\_rna\_seq$age2)

## Warning: Unknown or uninitialised column: `age2`.

## Warning in mean.default(tcga\_rna\_seq$age2): argument is not numeric or logical:  
## returning NA

## [1] NA

sd(tcga\_rna\_seq$age2)

## Warning: Unknown or uninitialised column: `age2`.

## [1] NA

bmi2=tcga\_rna\_seq %>%  
 group\_by(sample\_type)%>%  
 filter(bmi!="NA")%>%  
 summarize(mean=mean(bmi),  
 sd=sd(bmi))

## `summarise()` ungrouping output (override with `.groups` argument)

bmi2

## # A tibble: 3 x 3  
## sample\_type mean sd  
## <chr> <dbl> <dbl>  
## 1 Esophageal adeno-carcinoma tissues 28.8 6.84  
## 2 Esophageal adeno-carcinoma-adjacent tissues 36.2 12.6   
## 3 Esophageal squamous cell carcinoma tissue 21.8 2.83

mean(tcga\_rna\_seq$bmi)

## [1] NA

sd(tcga\_rna\_seq$bmi)

## [1] NA

tcga\_rna\_seq=tcga\_rna\_seq %>%  
 mutate(smoke=0,  
 smoke=ifelse(cigarettes\_per\_day>0,   
 "Yes", "No")  
 )  
  
tcga\_rna\_seq$barrettsid<-tcga\_rna\_seq$`Barrett's Esophagus Reported`  
  
barrettsid=tcga\_rna\_seq %>%  
 group\_by(sample\_type, barrettsid)%>%  
 summarize(N=n())%>%  
 ungroup(barrettsid)%>%  
 mutate(p=N/sum(N))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

barrettsid

## # A tibble: 7 x 4  
## # Groups: sample\_type [3]  
## sample\_type barrettsid N p  
## <chr> <chr> <int> <dbl>  
## 1 Esophageal adeno-carcinoma tissues No 17 0.607  
## 2 Esophageal adeno-carcinoma tissues Not Available 4 0.143  
## 3 Esophageal adeno-carcinoma tissues Yes 7 0.25   
## 4 Esophageal adeno-carcinoma-adjacent tissues No 5 0.714  
## 5 Esophageal adeno-carcinoma-adjacent tissues Yes 2 0.286  
## 6 Esophageal squamous cell carcinoma tissue No 13 0.419  
## 7 Esophageal squamous cell carcinoma tissue Not Available 18 0.581

smoke2=tcga\_rna\_seq %>%  
 group\_by(sample\_type, smoke)%>%  
 summarize(N=n())%>%  
 ungroup(smoke)%>%  
 mutate(p=N/sum(N))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

smoke2

## # A tibble: 6 x 4  
## # Groups: sample\_type [3]  
## sample\_type smoke N p  
## <chr> <chr> <int> <dbl>  
## 1 Esophageal adeno-carcinoma tissues Yes 19 0.679  
## 2 Esophageal adeno-carcinoma tissues <NA> 9 0.321  
## 3 Esophageal adeno-carcinoma-adjacent tissues Yes 4 0.571  
## 4 Esophageal adeno-carcinoma-adjacent tissues <NA> 3 0.429  
## 5 Esophageal squamous cell carcinoma tissue Yes 17 0.548  
## 6 Esophageal squamous cell carcinoma tissue <NA> 14 0.452

stage2=tcga\_rna\_seq %>%  
 group\_by(sample\_type, tumor\_stage)%>%  
 summarize(N=n())%>%  
 ungroup(tumor\_stage)%>%  
 mutate(p=N/sum(N))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

stage2

## # A tibble: 25 x 4  
## # Groups: sample\_type [3]  
## sample\_type tumor\_stage N p  
## <chr> <chr> <int> <dbl>  
## 1 Esophageal adeno-carcinoma tissues not reported 7 0.25   
## 2 Esophageal adeno-carcinoma tissues stage i 5 0.179   
## 3 Esophageal adeno-carcinoma tissues stage ia 1 0.0357  
## 4 Esophageal adeno-carcinoma tissues stage iia 2 0.0714  
## 5 Esophageal adeno-carcinoma tissues stage iib 5 0.179   
## 6 Esophageal adeno-carcinoma tissues stage iii 1 0.0357  
## 7 Esophageal adeno-carcinoma tissues stage iiia 1 0.0357  
## 8 Esophageal adeno-carcinoma tissues stage iiib 4 0.143   
## 9 Esophageal adeno-carcinoma tissues stage iiic 1 0.0357  
## 10 Esophageal adeno-carcinoma tissues stage iv 1 0.0357  
## # … with 15 more rows

survival2=tcga\_rna\_seq %>%  
 group\_by(sample\_type)%>%  
 filter(days\_to\_death!="NA")%>%  
 summarize(mean=mean(days\_to\_death),  
 sd=sd(days\_to\_death))

## `summarise()` ungrouping output (override with `.groups` argument)

survival2

## # A tibble: 3 x 3  
## sample\_type mean sd  
## <chr> <dbl> <dbl>  
## 1 Esophageal adeno-carcinoma tissues 444. 465.  
## 2 Esophageal adeno-carcinoma-adjacent tissues 313 423.  
## 3 Esophageal squamous cell carcinoma tissue 452 432.

# TCGA RNA WGS

tcga\_rna\_wgs <- microbiome.data.tcga.WGS$meta.dat  
  
as.data.frame(table(tcga\_rna\_wgs$morphology))

## Var1 Freq  
## 1 8070/3 55  
## 2 8071/3 7  
## 3 8083/3 2  
## 4 8140/3 29  
## 5 8480/3 2

tcga\_rna\_wgs=tcga\_rna\_wgs %>%  
 mutate(sample\_type=0,  
 sample\_type=ifelse(morphology=="8070/3",   
 "Squamous cell carcinoma, NOS",   
 sample\_type),  
 sample\_type=ifelse(morphology=="8071/3",   
 "Squamous cell carcinoma, keratinizing, NOS",   
 sample\_type),  
 sample\_type=ifelse(morphology=="8083/3",   
 "Basaloid squamous cell carcinoma",   
 sample\_type),  
 sample\_type=ifelse(morphology=="8140/3",   
 "Adenocarcinoma, NOS", sample\_type),  
 sample\_type=ifelse(morphology=="8480/3",   
 "Mucinous adenocarcinoma",   
 sample\_type)  
 )  
table(tcga\_rna\_wgs$sample\_type)

##   
## Adenocarcinoma, NOS   
## 29   
## Basaloid squamous cell carcinoma   
## 2   
## Mucinous adenocarcinoma   
## 2   
## Squamous cell carcinoma, keratinizing, NOS   
## 7   
## Squamous cell carcinoma, NOS   
## 55

N3=table(tcga\_rna\_wgs$sample\_type)  
  
sex3=tcga\_rna\_wgs %>%  
 group\_by(sample\_type, Gender)%>%  
 summarize(N3=n())%>%  
 ungroup(Gender)%>%  
 mutate(p=N3/sum(N3))%>%  
 ungroup(sample\_type)%>%  
 group\_by(Gender)%>%  
 mutate(t=sum(N3))%>%  
 ungroup(Gender)%>%  
 mutate(pt=t/sum(N3))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

sex3

## # A tibble: 9 x 6  
## sample\_type Gender N3 p t pt  
## <chr> <chr> <int> <dbl> <int> <dbl>  
## 1 Adenocarcinoma, NOS female 9 0.310 18 0.146  
## 2 Adenocarcinoma, NOS male 20 0.690 77 0.626  
## 3 Basaloid squamous cell carcinoma female 2 1 18 0.146  
## 4 Mucinous adenocarcinoma male 2 1 77 0.626  
## 5 Squamous cell carcinoma, keratinizing, NOS female 3 0.429 18 0.146  
## 6 Squamous cell carcinoma, keratinizing, NOS male 4 0.571 77 0.626  
## 7 Squamous cell carcinoma, NOS female 4 0.0727 18 0.146  
## 8 Squamous cell carcinoma, NOS male 51 0.927 77 0.626  
## 9 <NA> <NA> 28 1 28 0.228

race3=tcga\_rna\_wgs %>%  
 group\_by(sample\_type, race)%>%  
 summarize(N3=n())%>%  
 ungroup(race)%>%  
 mutate(p=N3/sum(N3))%>%  
 ungroup(sample\_type)%>%  
 group\_by(race)%>%  
 mutate(t=sum(N3))%>%  
 ungroup(race)%>%  
 mutate(pt=t/sum(N3))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

race3

## # A tibble: 9 x 6  
## sample\_type race N3 p t pt  
## <chr> <chr> <int> <dbl> <int> <dbl>  
## 1 Adenocarcinoma, NOS asian 1 0.0345 37 0.301   
## 2 Adenocarcinoma, NOS white 28 0.966 55 0.447   
## 3 Basaloid squamous cell carcinoma white 2 1 55 0.447   
## 4 Mucinous adenocarcinoma white 2 1 55 0.447   
## 5 Squamous cell carcinoma, keratin… white 7 1 55 0.447   
## 6 Squamous cell carcinoma, NOS asian 36 0.655 37 0.301   
## 7 Squamous cell carcinoma, NOS black or african … 3 0.0545 3 0.0244  
## 8 Squamous cell carcinoma, NOS white 16 0.291 55 0.447   
## 9 <NA> <NA> 28 1 28 0.228

age3=tcga\_rna\_wgs %>%  
 group\_by(sample\_type)%>%  
 filter(year\_of\_birth!="NA")%>%  
 summarize(mean=mean(year\_of\_birth),  
 sd=sd(year\_of\_birth))

## `summarise()` ungrouping output (override with `.groups` argument)

age3

## # A tibble: 5 x 3  
## sample\_type mean sd  
## <chr> <dbl> <dbl>  
## 1 Adenocarcinoma, NOS 1939. 9.89  
## 2 Basaloid squamous cell carcinoma 1956 0   
## 3 Mucinous adenocarcinoma 1928 0   
## 4 Squamous cell carcinoma, keratinizing, NOS 1953. 9.27  
## 5 Squamous cell carcinoma, NOS 1952. 11.6

mean(tcga\_rna\_wgs$year\_of\_birth)

## [1] NA

sd(tcga\_rna\_wgs$year\_of\_birth)

## [1] NA

bmi3=tcga\_rna\_wgs %>%  
 group\_by(sample\_type)%>%  
 filter(bmi!="NA")%>%  
 summarize(mean=mean(bmi),  
 sd=sd(bmi))

## `summarise()` ungrouping output (override with `.groups` argument)

bmi3

## # A tibble: 4 x 3  
## sample\_type mean sd  
## <chr> <dbl> <dbl>  
## 1 Adenocarcinoma, NOS 33.6 11.8   
## 2 Mucinous adenocarcinoma 39.5 0   
## 3 Squamous cell carcinoma, keratinizing, NOS 21.1 3.49  
## 4 Squamous cell carcinoma, NOS 22.1 2.97

mean(tcga\_rna\_wgs$bmi)

## [1] NA

sd(tcga\_rna\_wgs$bmi)

## [1] NA

tcga\_rna\_wgs=tcga\_rna\_wgs %>%  
 mutate(smoke=0,  
 smoke=ifelse(cigarettes\_per\_day>0,   
 "Yes", "No")  
 )  
  
smoke3=tcga\_rna\_wgs %>%  
 group\_by(sample\_type, smoke)%>%  
 summarize(N3=n())%>%  
 ungroup(smoke)%>%  
 mutate(p=N3/sum(N3))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

smoke3

## # A tibble: 9 x 4  
## # Groups: sample\_type [6]  
## sample\_type smoke N3 p  
## <chr> <chr> <int> <dbl>  
## 1 Adenocarcinoma, NOS Yes 18 0.621  
## 2 Adenocarcinoma, NOS <NA> 11 0.379  
## 3 Basaloid squamous cell carcinoma Yes 2 1   
## 4 Mucinous adenocarcinoma Yes 2 1   
## 5 Squamous cell carcinoma, keratinizing, NOS Yes 4 0.571  
## 6 Squamous cell carcinoma, keratinizing, NOS <NA> 3 0.429  
## 7 Squamous cell carcinoma, NOS Yes 23 0.418  
## 8 Squamous cell carcinoma, NOS <NA> 32 0.582  
## 9 <NA> <NA> 28 1

stage3=tcga\_rna\_wgs %>%  
 group\_by(sample\_type, tumor\_stage)%>%  
 summarize(N3=n())%>%  
 ungroup(tumor\_stage)%>%  
 mutate(p=N3/sum(N3))

## `summarise()` regrouping output by 'sample\_type' (override with `.groups` argument)

stage3

## # A tibble: 18 x 4  
## # Groups: sample\_type [6]  
## sample\_type tumor\_stage N3 p  
## <chr> <chr> <int> <dbl>  
## 1 Adenocarcinoma, NOS not reported 2 0.0690  
## 2 Adenocarcinoma, NOS stage i 10 0.345   
## 3 Adenocarcinoma, NOS stage ia 2 0.0690  
## 4 Adenocarcinoma, NOS stage iib 8 0.276   
## 5 Adenocarcinoma, NOS stage iiib 3 0.103   
## 6 Adenocarcinoma, NOS stage iiic 2 0.0690  
## 7 Adenocarcinoma, NOS stage iv 2 0.0690  
## 8 Basaloid squamous cell carcinoma stage ia 2 1   
## 9 Mucinous adenocarcinoma not reported 2 1   
## 10 Squamous cell carcinoma, keratinizing, NOS stage ib 2 0.286   
## 11 Squamous cell carcinoma, keratinizing, NOS stage iia 5 0.714   
## 12 Squamous cell carcinoma, NOS not reported 3 0.0545  
## 13 Squamous cell carcinoma, NOS stage iia 22 0.4   
## 14 Squamous cell carcinoma, NOS stage iib 8 0.145   
## 15 Squamous cell carcinoma, NOS stage iii 9 0.164   
## 16 Squamous cell carcinoma, NOS stage iiia 9 0.164   
## 17 Squamous cell carcinoma, NOS stage iv 4 0.0727  
## 18 <NA> <NA> 28 1