Rat DCIS study

Anne Trinh

2020-11-07

Contents

1 Prerequisites **5**
- 1.1 Packages and Software . . . 5
- 1.2 External software . . . 6
- 1.3 Annotations . . . 7

2 Cohort summary **11**
- 2.1 Calculating growth rates . . . 11
- 2.2 FACS data . . . 17
- 2.3 Summary Table . . . 20

3 Whole-slide imaging **23**
- 3.1 Associate the frequencies with other data types . . . 23
- 3.2 Cellular composition . . . 25
- 3.3 Associate composition with other covariates . . . 29
- 3.4 Estimate tumor size . . . 32
- 3.5 Correlations between different subpopulations . . . 32
- 3.6 Associations between CD8 counts with other clinical variables . . . 34

4 Spatial statistics **37**
- 4.1 knn-Distances: . . . 37
- 4.2 The interacting fraction . . . 44
- 4.3 M-H distances . . . 50
- 4.4 Comparison between metrics . . . 57
- 4.5 Other cell types . . . 61

5 Expression data **65**
- 5.1 Running alignment . . . 65
- 5.2 RNA Initial QC . . . 66
- 5.3 Normalisation . . . 69
- 5.4 Processing files for external software . . . 72

6 RNA data: preliminary plots **73**
- 6.1 PCA plots . . . 73
- 6.2 Expression patterns by cell type . . . 79

7 DESeq analysis **83**
7.1 DN vs Ep 83
7.2 Set-up the comparisons 86
7.3 PCA plots 89
7.4 GSEA 98
7.5 CD45 comparisons 99
7.6 DN comparisons 99

8 DESeq analysis **101**
8.1 Focus mainly on growing vs stable: 102

This is a summary of the code required in the study: 'insert study name'

Chapter 1

Prerequisites

1.1 Packages and Software

The following packages are required to conduct the analyses described below.

In house scripts are deposited in the rscript folder.

```
library(AnnotationHub)
library(biomaRt)
library(Biostrings)
library(colorspace)
library(DESeq2)
library(dplyr)
library(DT)
library(ensembldb)
library(EnsDb.Hsapiens.v86)
library(GenVisR)
library(GenomicFeatures)
library(ggplot2)
library(gplots)
library(GSEABase)
library(GSVA)
library(heatmap.plus)
library(HTSanalyzeR2)
#library(kableExtra)
library(limma)
library(matrixStats)
library(pamr)
library(reshape2)
library(RColorBrewer)
library(scales)
```

```
library(spatstat)
library(tcR)
library(vcfR)
library(xlsx)
library(writexl)

DiffCols=hue_pal()(8)
palette(brewer.pal(9, "Set1"))
RdBu=brewer.pal(11, "RdBu")
SetCols=brewer.pal(12, "Set3")

source("../rscript/cnFreq_fn.R") #modified version of GenVisR
source("../rscript/merge_contig.R")
source("../rscript/gseaCode.R")
source("../rscript/ContingencyTable.R")
source("../rscript/PvalueHeatMap.R")
source("../rscript/BootstrapShannonIdx.R")
source("../rscript/CreateRnor87db.R")
source("../rscript/FindRatAAHomolog.R")

firstup <- function(x) {
  substr(x, 1, 1) <- toupper(substr(x, 1, 1))
  x
}

ColMerge=matrix(c("#FFC82F", "#FFEDBC", "#73FDFE","#D2FFFF", "#FF41FF", "#FECAFF", "#5
rownames(ColMerge)=c("LY","PDL1", "PDL1+LY","Vehicle")

Hsedb<-EnsDb.Hsapiens.v86
```

1.2 External software

The following external software was utilised:

Software	Function
bwa	Alignment of WGS data to reference
GATK4	Mutation calling, done by NYGC. Mutation calling from RNA (Haplotype caller)
strelka	Mutation calling, done by NYGC
BICseq	CNV calling
GEM3	create mappability files for CNV calling
STAR	Alignment of RNAseq data
RSEM	Calculate RSEM, TPM, FPKM from RNAseq data
TRUST4	assignment of T and B cell clonotypes from RNA-seq data

Software	Function
Oncotator	Annotation of genetic variants
QuPath	Tool for cell segmentation and extraction of features from IF images
samtools, bcftools	querying and dispalying information from bam files, extracting allelic depth at specific genomic locations
CIBERSORT	Inferring immune composition from RNA
lumpy	structural variants
PAM50	code from parker et al 2009 to infer PAM50 subtypes

1.3 Annotations

1.3.1 Genomic properties

Information on chromosome sizes, cytobands and centromere locations were obtained from the UCSC genome browser.

The following annotation data for the hg19/b37/GRCh37 genome is required:

Data Type	Download link
ref. genome	http://hgdownload.soe.ucsc.edu/goldenPath/rn6/bigZips/rn6.fa.gz
refSeq annot	http://hgdownload.soe.ucsc.edu/goldenPath/rn6/bigZips/genes/rn6.refGene.gtf.gz
refSeq annot	http://hgdownload.soe.ucsc.edu/goldenPath/rn6/bigZips/genes/rn6.ncbiRefSeq.gtf.gz
gff3 file	for TRUST4 (ftp://ftp.ensembl.org/pub/release-100/gff3/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.100.gff3.gz)
gene lengths	Extracted from GRCh37.75.gtf file from ensembl
hg19cytoBand	ucsc server of all cytoband locations
chromosome sizes	uscs genome browser
biomart	conversion between gene symbol, ensbl and entrez was faciliated using biomart package

Below is the summary of chromosome sizes and centromere locations:

```
##   chrom chromStart chromEnd name gieStain
## 1  chr1          0 10704345  p13     gneg
## 2  chr1   10704345 25618263  p12     gvar
## 3  chr1   25618263 40652454  p11     gneg
## 4  chr1   40652454 51356799  q11     gpos
## 5  chr1   51356799 73607402  q12     gneg
## 6  chr1   73607402 95016091  q21     gpos
```

Create a TxDb object from a gtf file and save information: Not sure which version to use??

Below is an example of the gene annotation files

```
## GRanges object with 6 ranges and 2 metadata columns:
##          seqnames          ranges strand |     gene_id gene_width
##             <Rle>       <IRanges>  <Rle> | <character>  <integer>
##   Vom2r3     chr1   396700-409676      + |      Vom2r3      12977
##    Lrp11     chr1 1702696-1731210      + |       Lrp11      28515
##    Nup43     chr1 1771721-1781554      + |       Nup43       9834
##    Lats1     chr1 1784078-1817310      + |       Lats1      33233
##   Katna1     chr1 1826170-1867786      + |      Katna1      41617
##    Ppil4     chr1 1897350-1930311      + |       Ppil4      32962
##   -------
##   seqinfo: 22 sequences from an unspecified genome; no seqlengths
```

1.3.2 Gene name homologs between organisms

Biomart was used to convert between rat, mouse and human gene symbols and ensembl ids. Below is an example of the human gene names mapped to the rat homolog

Figure out what is required here for later analysis

```
##   HGNC.symbol RGD.symbol
## 1        TAB1       Tab1
## 2        PHF1       Phf1
## 3       RNF39      Rnf39
## 4      IGSF10     Igsf10
## 5     TMEM130    Tmem130
## 6       EFNB1      Efnb1
```

1.3.3 Gene signatures and data-bases

Gene sets/signatures were obtained from the following sources:

Source	Description
IEDB	database of immune epitopes
MsigDB	c2, c5, hallmark set of curated pathway gene sets
Metacore	Process Networks and Pathway Maps data bases
COSMIC	database of concensus oncogenes
ImmPort	List of immune related genes
InnateDB	List of genes associated with innate immune system
Rosenthal 2019	genes associated with MHC-I presentation
Thorsson 2018	Immune gene signatures curated from studies by Wolf, Calabro, Teschendorff, Beck, Chang

Source	Description
Pardoll, Wykes	Immune checkpoint genes
gil del alcazar 2017	Supplementary table 5: list of activation, dysfunction gene signatures
Bailey 2018	List of 10 most comon tumor pathways
Chang 2018	Common mutation locations in cancer

The PAM50 signature was implemented using the scripts provided in supplementary from Parker et al 2010. These gene lists can be found in the annotations folder

1.3.3.1 Human gene homologs

Below, lists of common mutations in cancer are loaded and the "homolog" in rat is determined using an in-house script. The steps involved are:

- determine the amino acid context in human (find 5 a.a. prior and after)
- find the region with most amino acid homology in rat (2 or less differences)
- check whether the amino acid of interest is present in rat

An example of the output is shown

```
##   Gene AAno AA1 Variant RatGene    Sequence         HumProt             RatProt
## 1 NRAS   61   Q       R    Nras LDTAGQEEYSA ENSP00000358548 ENSRNOP00000036381
## 2 NRAS   61   Q       K    Nras LDTAGQEEYSA ENSP00000358548 ENSRNOP00000036381
## 3 NRAS   61   Q       L    Nras LDTAGQEEYSA ENSP00000358548 ENSRNOP00000036381
## 4 NRAS   61   Q       H    Nras LDTAGQEEYSA ENSP00000358548 ENSRNOP00000036381
## 5 NRAS   61   Q       P    Nras LDTAGQEEYSA ENSP00000358548 ENSRNOP00000036381
## 6 NRAS   61   Q       *    Nras LDTAGQEEYSA ENSP00000358548 ENSRNOP00000036381
##   RatAAno RatSequence
## 1      61 LDTAGQEEYSA
## 2      61 LDTAGQEEYSA
## 3      61 LDTAGQEEYSA
## 4      61 LDTAGQEEYSA
## 5      61 LDTAGQEEYSA
## 6      61 LDTAGQEEYSA
```

1.3.3.2 GSEA compendiums

For pathway analysis, the c2 (pathway), Hallmark and c5 (Gene Ontology). In addition, metacore pathways (pathway maps and process networks) were obtained and loaded below. This gives a list of 7 different data-sets to interrogate CHECK: When GSVA is run, which input is required??

Chapter 2

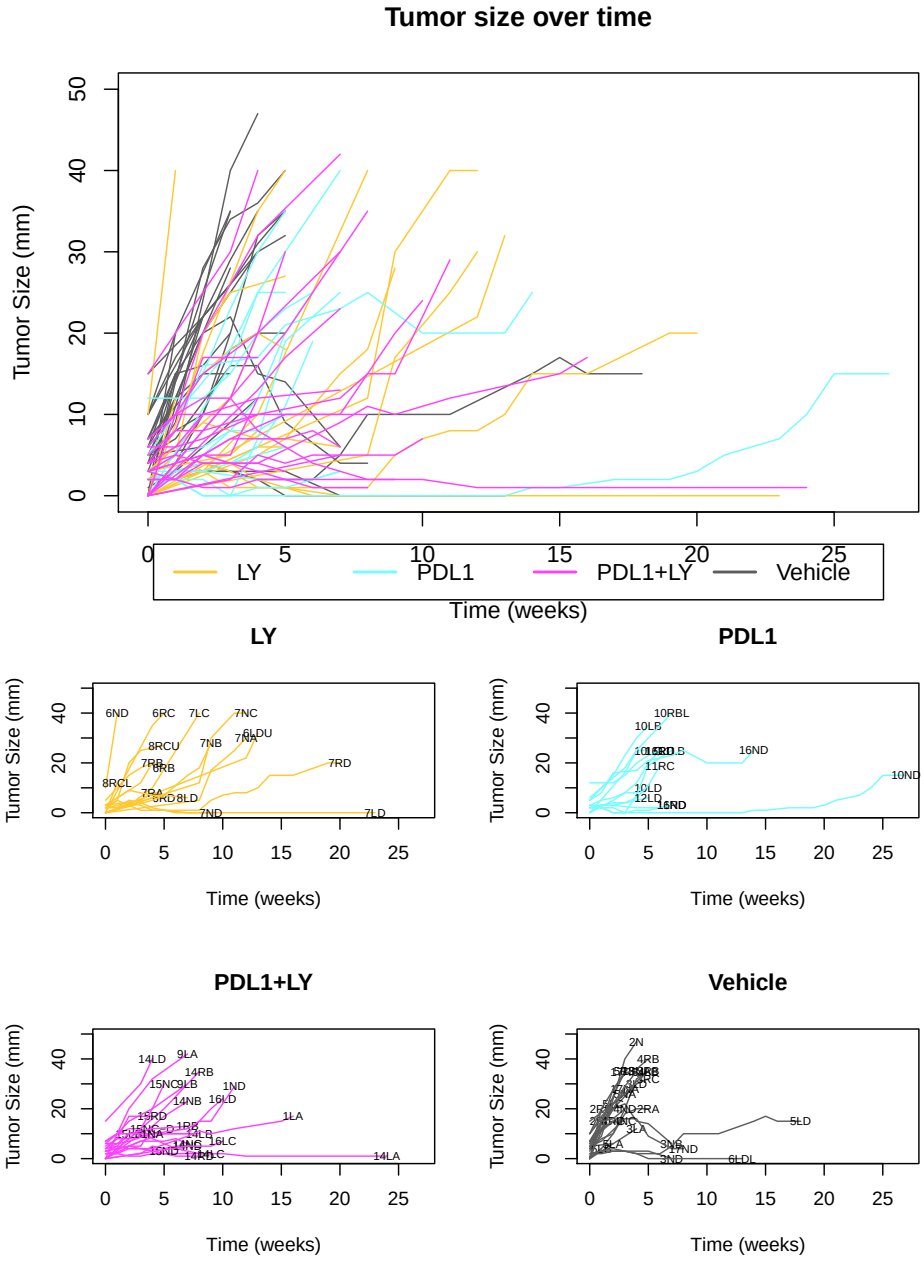
Cohort summary

Below we assess summary statistics on clinico-pathological features of this data set. This includes information on:

- treatment
- tumor size
- growth rates (mm/week)
- number of tumors per rat

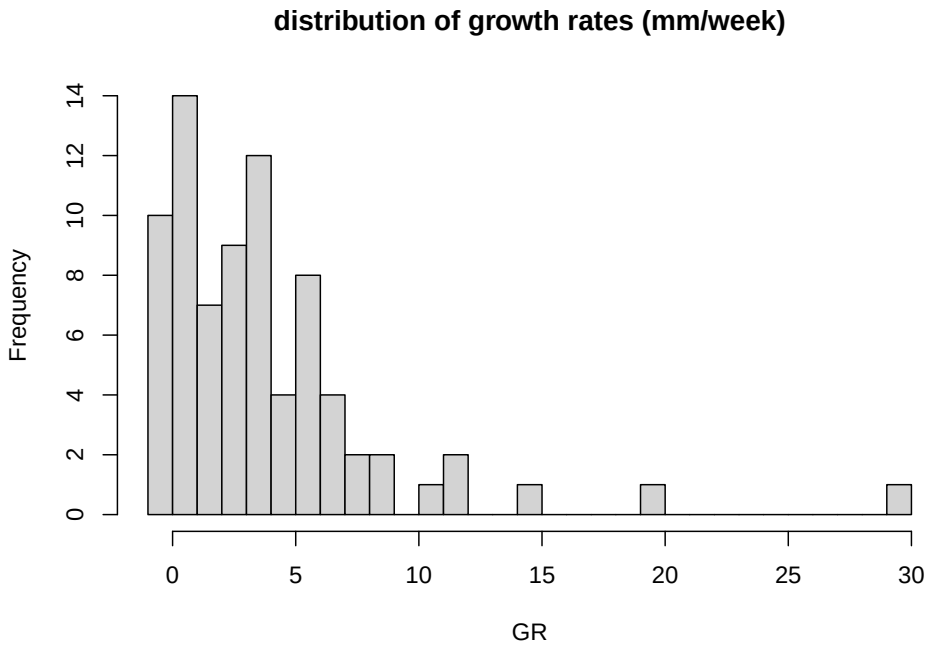
2.1 Calculating growth rates

In this section, we estimate the growth rates of the samples: Below is a plot of the tumor size per week for each recorded tumor, color-coded according to treatment. Time is measured at the first time point at which a tumor is palpated. Spontaneous large tumors are assumed to have a tumor size of 0 or 1 one week prior to palpating.

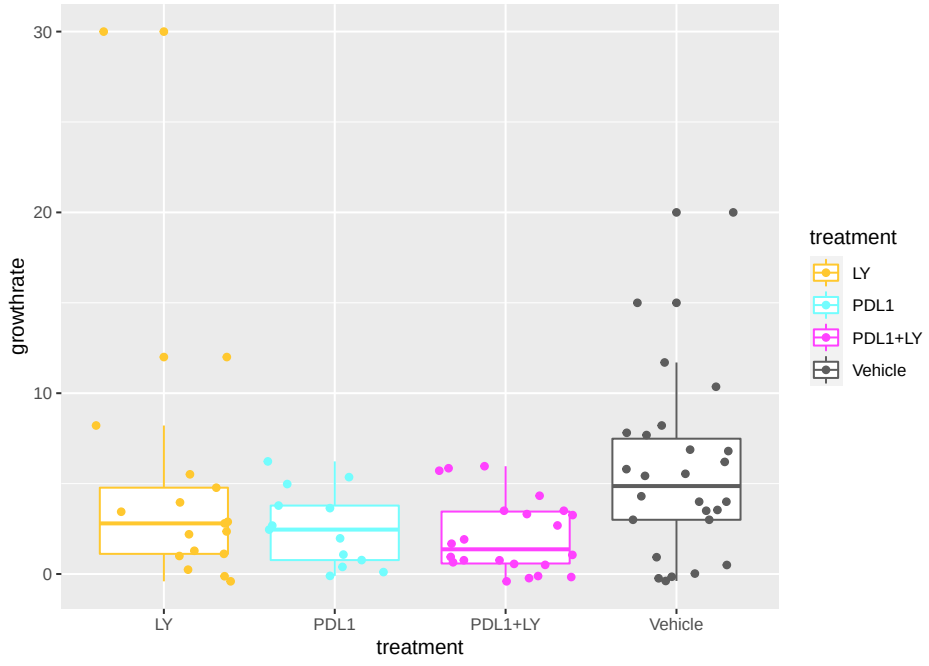


These tumors doesn't have growth rate data: 17N_B.

We can then compute the growth rate for the above samples by considering the change in size over a given period of time using a linear regression model. Below is the histogram of growth rates:



Based on the above distribution, a cut-off of 2mm/week may be an optimal cut-off to separate growing and stable tumors. Below are growth rates of tumors under different treatments:



We can calculate the p.values below, using a t.test. The growth rates comparing the treatment to the controls are:

```
## [1] "LY samples"
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d1$growthrate[d1$treatment == "LY"] and d1$growthrate[d1$treatment == "Vehicl
## W = 165.5, p-value = 0.1718
## alternative hypothesis: true location shift is not equal to 0
```

```
## [1] "PDL1 samples"
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d1$growthrate[d1$treatment == "PDL1"] and d1$growthrate[d1$treatment == "Vehi
## W = 103, p-value = 0.051
## alternative hypothesis: true location shift is not equal to 0
```

```
## [1] "PDL1+LY samples"
```

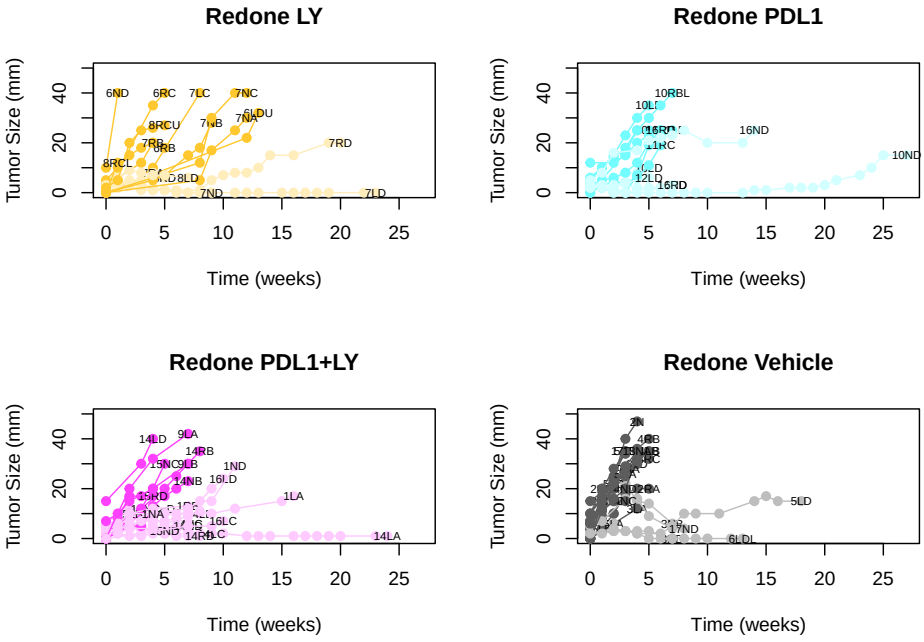
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d1$growthrate[d1$treatment == "PDL1+LY"] and d1$growthrate[d1$treatment == "V
## W = 154.5, p-value = 0.006713
## alternative hypothesis: true location shift is not equal to 0
```

This shows a smaller growth-rate in PDL1 single and double treated cases compared to the vehicles.

Overall the distribution of growing vs stable tumors is shown below:

```
##
##   grow stable
##     47     31
```

We can replot the previous graphs according to growth, and color code according to whether it is a fast or slow growing tumor



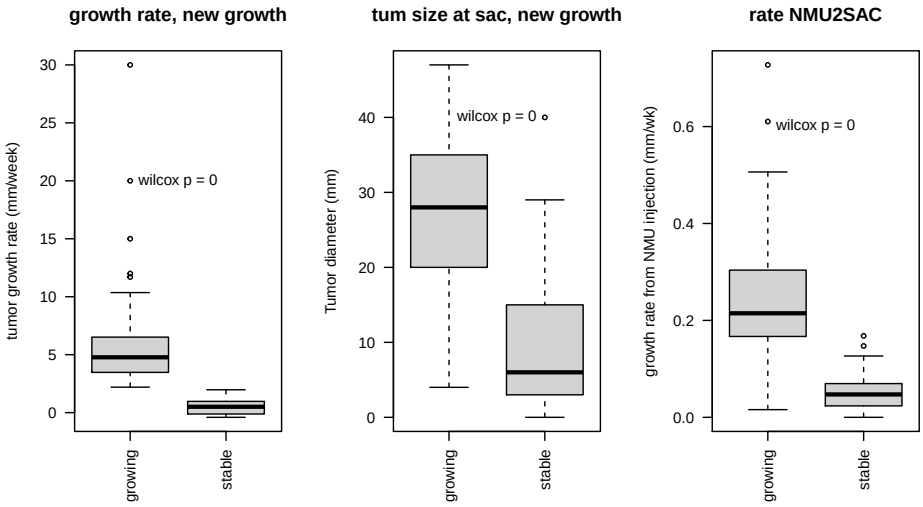
As a sanity check, compare these growth rates with differences in tumour size at different time points:

- comparing the growth rate according to classifications (growing, stable)
- tumor size at time of sacrifice
- rate of tumor development from the time of NMU injection

```
## Warning in wilcox.test.default(x = c(3.54285714285714, 15, 20, 11.7, 5.8, :
## cannot compute exact p-value with ties
```

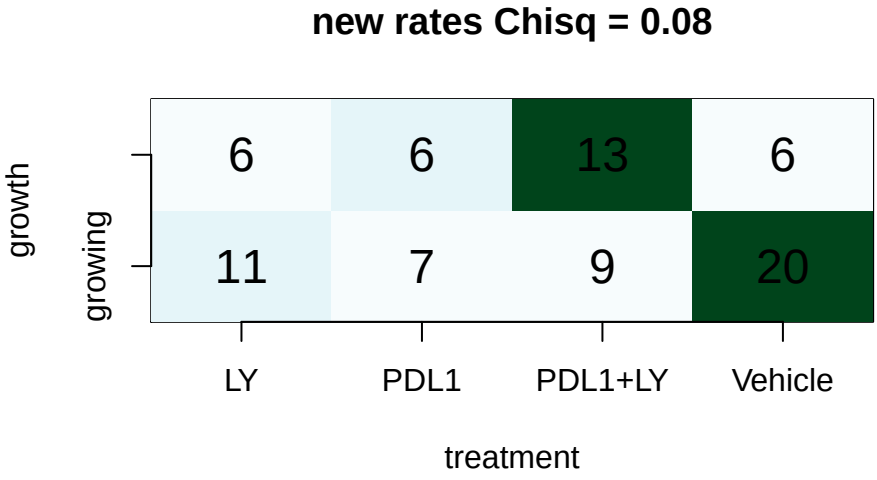
```
## Warning in wilcox.test.default(x = c(20, 15, 20, 47, 30, 12, 32, 35, 4, : cannot
## compute exact p-value with ties
```

```
## Warning in wilcox.test.default(x = c(0.238095238095238, 0.178571428571429, :
## cannot compute exact p-value with ties
```



Is there an association with treatment? Calculate below using chi-squared test:

```
##
##  Pearson's Chi-squared test
##
## data:  table(factor(d1$treatment), d1$growthrate_cutoff2)
## X-squared = 6.8181, df = 3, p-value = 0.07793
```



Overall, it appears that there is an association between growth rate and treatment

2.2 FACS data

The immune (CD45) fractions from a number of samples were collected, and assessed using FACs. The major cell types detected are:

Leukocytes:

- Tregs
- CD8 T cells
- Thelper cells
- B cells
- NK T cells
- gamma delta T cells

Myeloid cells:

- Macrophages M1
- Macrophages M2
- Dendritic cells
- Monocytes
- Neutrophils

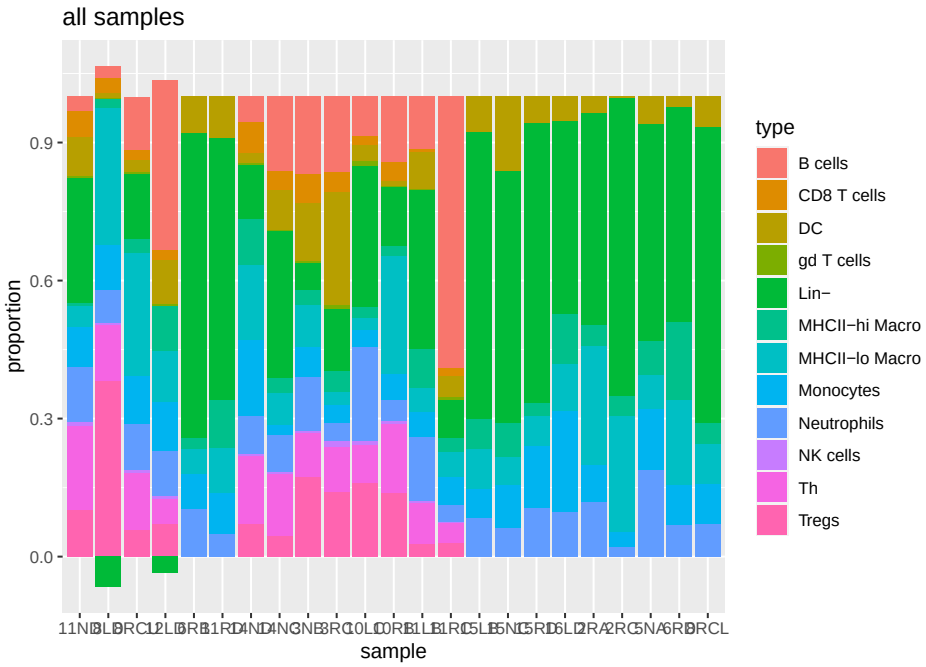
We can look at the:

- types of cells
- distributions

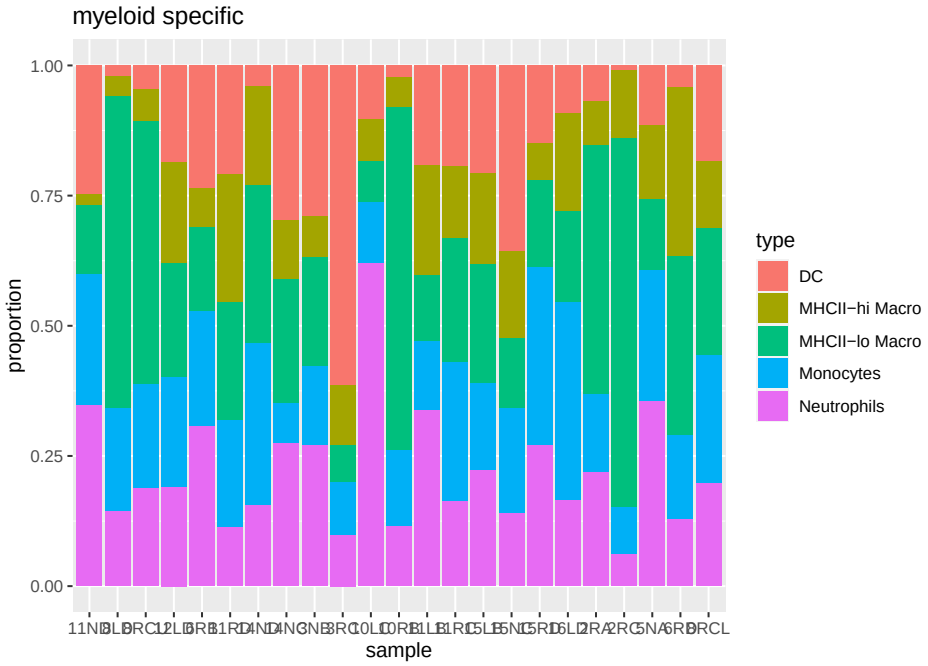
Note that in a number of samples the leukocyte population could not be inferred with confidence, and proportions are normalised to the myeloid population

```
##          type         11ND          8LD         8RCU         12LD 6RB 11RD
## 1  Leukocytes 913.84000000 379.56000000 829.57000000 298.83000000  NA   NA
## 2  leukocytes           NA           NA           NA           NA  NA   NA
## 3                       NA           NA           NA           NA  NA   NA
## 4          Th   0.18237328   0.12232585   0.12397989   0.05427835  NA   NA
## 5       Tregs   0.10207476   0.38225840   0.05701749   0.07030753  NA   NA
## 6 CD8 T cells   0.05586317   0.03293287   0.02488036   0.02399357  NA   NA
##           14ND         14NC          3NB          3RC         10LC         10RB
## 1 291.43000000 576.74000000 1.118820e+03 616.38000000 414.75000000 899.60000000
## 2           NA           NA           NA           NA           NA           NA
## 3           NA           NA           NA           NA           NA           NA
## 4   0.15032769   0.13704616 9.641408e-02   0.09800772   0.08323086   0.14899956
## 5   0.07030848   0.04386725 1.722261e-01   0.13929719   0.16072333   0.13873944
## 6   0.06811241   0.04036481 6.289662e-02   0.04382037   0.02013261   0.04357492
##           11LB         11RC 15LB 15NC 15RD 16LD 2RA 2RC 5NA 6RD 8RCL
## 1 1.243790e+03 204.89000000   NA   NA   NA   NA  NA  NA  NA  NA   NA
## 2           NA           NA   NA   NA   NA   NA  NA  NA  NA  NA   NA
## 3           NA           NA   NA   NA   NA   NA  NA  NA  NA  NA   NA
## 4 9.106039e-02   0.04543902   NA   NA   NA   NA  NA  NA  NA  NA   NA
## 5 2.633885e-02   0.02923520   NA   NA   NA   NA  NA  NA  NA  NA   NA
```

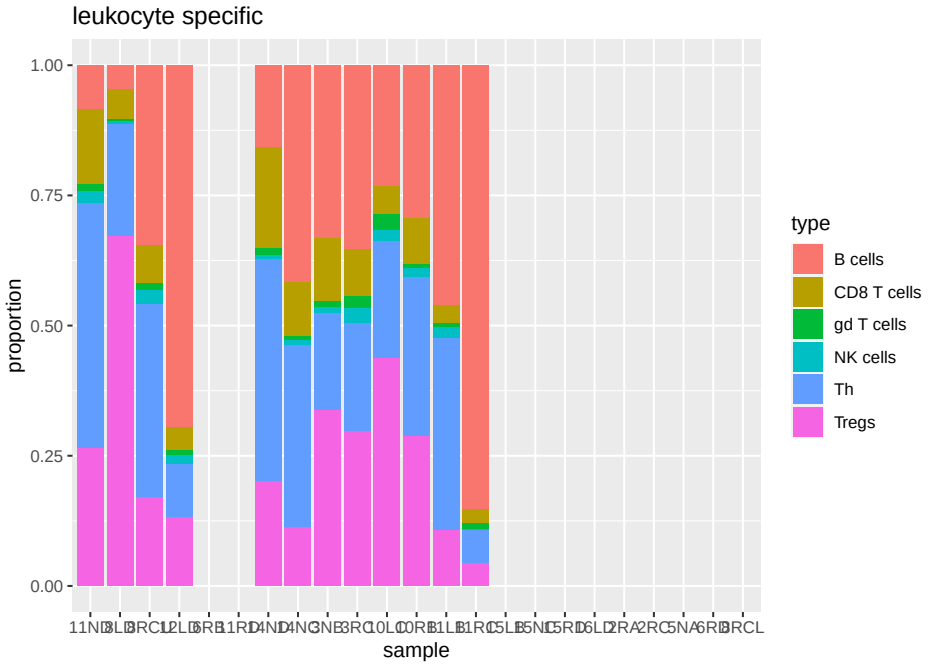
```
## 6 8.868057e-03   0.01883938   NA   NA   NA   NA   NA   NA   NA   NA   NA
```



We can look solely at the myeloid population (and normalise to this total), and color according to growth



Similarly, we can look at the leukocyte population. Note that the Treg population in some of these samples is very high.



2.3 Summary Table

Firstly, look at the total number of samples:

Feature	Levels	N
Total No Tumors	All	143
-	Characterisation	33
-	Prevention	8
-	Progression	84
Total No Rats	All	69
-	Characterisation	18
-	Prevention	8
-	Progression	42
No tumours per rat		1.84 (1, 3)

2.3.1 Compare the characterisation vs progression cohort

Feature	Levels	Characterisation	Progression
Total No tumours		59	102
Treatments	Vehicle	64	27
-	Untreated	4	
-	LY		17
-	PDL1		17
-	PDL1+LY		23
-	NA	18	18
Histology	diff. adenocarcinomas	41	76
-	mucinous carcinoma	0	3
-	Fibroadenoma	0	4
-	NA	18	19
Age Injection	32-36	8	84
-	35	27	0
-	49	6	0
-	NA	18	18
Time (days)	NMU 2 Sac	98.59 (54, 160)	149.37 (79, 248.7)
-	Cases with NA	30	18
-	NMU 2 Tumor		100.62 (44.2, 182.7)
-	Cases with NA	59	18
-	Tum Spec Surv		48.75 (9.5, 92)
-	Cases with NA	59	18
Growth Rate/Size (mm)	overall size @ sac	6.67 (1.4, 12.4)	19.52 (3, 40)
-	Growing No.		47
-	Growing size @ sac		27.7 15, 28, 40
-	Stable		31
-	Stable size @ sac		9.58 (1, 6, 24)

Feature	Levels	Characterisation	Progression
-	NA	59	24
Spatial Pattern			
-	Infiltrating	0	26
-	Restricted	0	33
-	NA	59	43
RNA samples	any fraction	14	46
-	Ep	14	21
-	DN	0	33
-	CD45	10	36
Imaging Data	No tumors		64
-	No tumors with RNA		38
FACS data	Comprehensive		22
-	EpCAM/CD45	0	82

TO INCLUDE: - trichrome data?

2.3.2 Summary of the RNA data

Below is a table of the samples with RNA information

Feature	Levels	Characterisation	Progre
RNA samples	any fraction	14	46
-	Ep	14	21
-	DN	0	33
-	CD45	10	36
Treatments Char/Prev	Vehicle	13	15
-	Untreat (char)	1	
-	LY		9
-	PDL1		11
-	PDL1+LY		11
-	NA	0	0
Time	NMU 2 Sac	106.9 (54.9, 160)	113.85
Growth Rate/Size (mm)	overall size @ sac	7.1 (3.9, 12.2)	21.96
-	Growing No.		33
-	Growing size @ sac		27.64
-	Stable		12
-	Stable size @ sac		7.17 (3
-	NA	14	1
Growth and Treatment: comparing small/stable vs large/growing			
Vehicle	N s/l		2, 13
LY	N s/l		2, 7
PDL1 and Treatment	N s/l		4, 6
PDL1+LY	N s/l		4, 7

Chapter 3

Whole-slide imaging

In this section, we will be looking at the composition and spatial distribution of cells in whole slide images. These sections have previously been assessed using an external script. Save this somewhere

The following markers have been used:

- EpCAM (tumor cells)
- SMA (fibroblasts or myeopithelial cells)
- CD8 (T cells)

Note that in some images a double positive EpCAM+/SMA+ population exists. Some CD8 cells have Epcam+ or SMA+ staining, however, we consider all of these to be simply CD8+

3.1 Associate the frequencies with other data types

Note there are 47 samples with imaging data. 33 of these samples have FACS data, manual counts and TIMER scores

Correlate the following information:

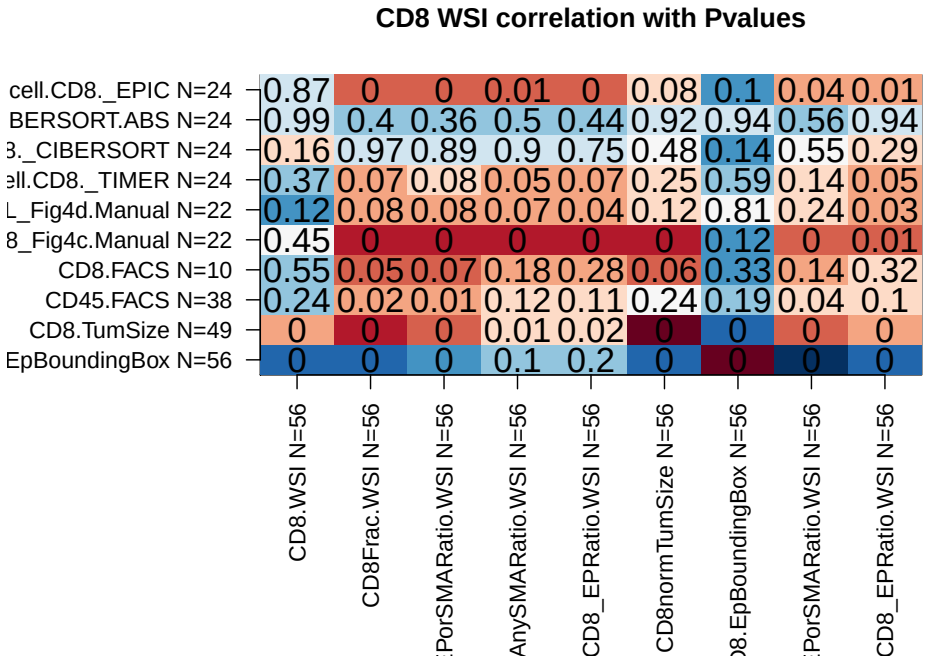
- "CD8.WSI": CD8 total counts
- "CD8Frac.WSI": CD8 fraction (normalised by cell count)
- "CD8_EPorSMARatio.WSI": CD8/EP+SMA ratio (any EpCAM or SMA + cell)
- "CD8_AnySMARatio.WSI": CD8/Any EPcam+ cell
- "CD8_EPRatio.WSI": CD8 to EpCAM+SMA- ratio
- "CD8normTumSize": normalised CD8 counts per mm of tumor size at sac
- "CD8.EpBoundingBox": approx area per CD8 cell (density)

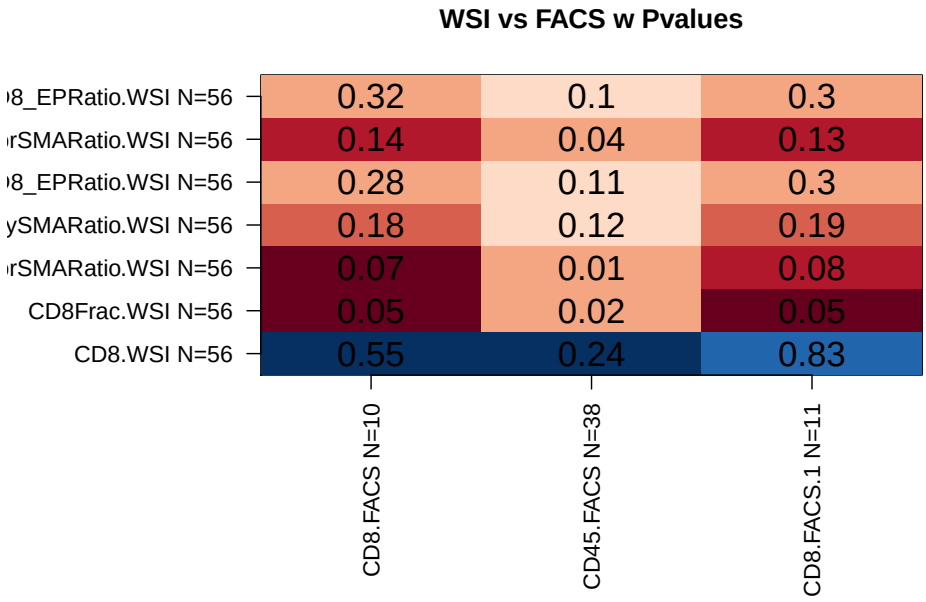
UPDATE THE CIBERSORT INFORMATION.

Below are heatmaps which show the correlation between two variables (red is correlated and blue is anti-correlated), and the p.value is indicated in the middle of the square.

It appears that CD8 whole-slide imaging associates well with:

- FACS data (both CD8 and CD45)
- Manual scoring (Fig 4) of CD8 cells
- Some CD8 gene signature scores (mainly in EPC, TIMER)



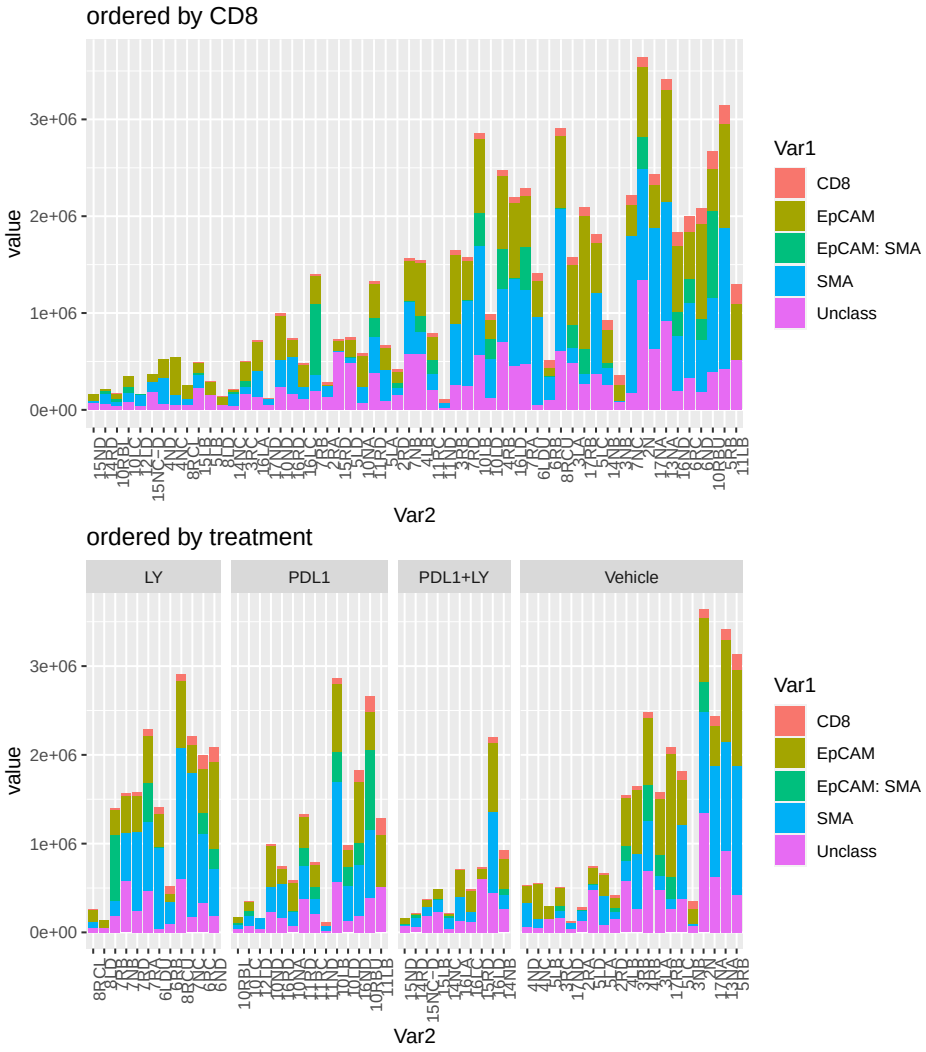


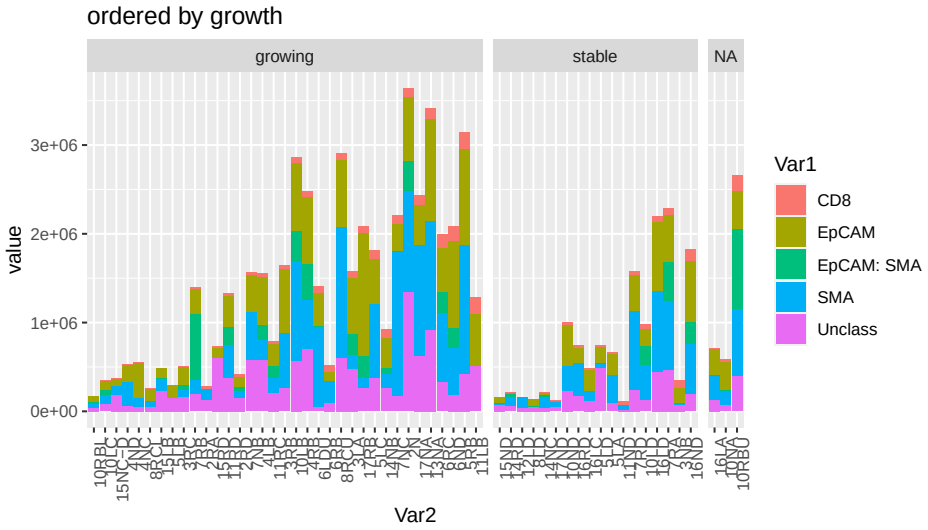
3.2 Cellular composition

Here, we look at the raw distributions of the different cell types and see if there are associations with:

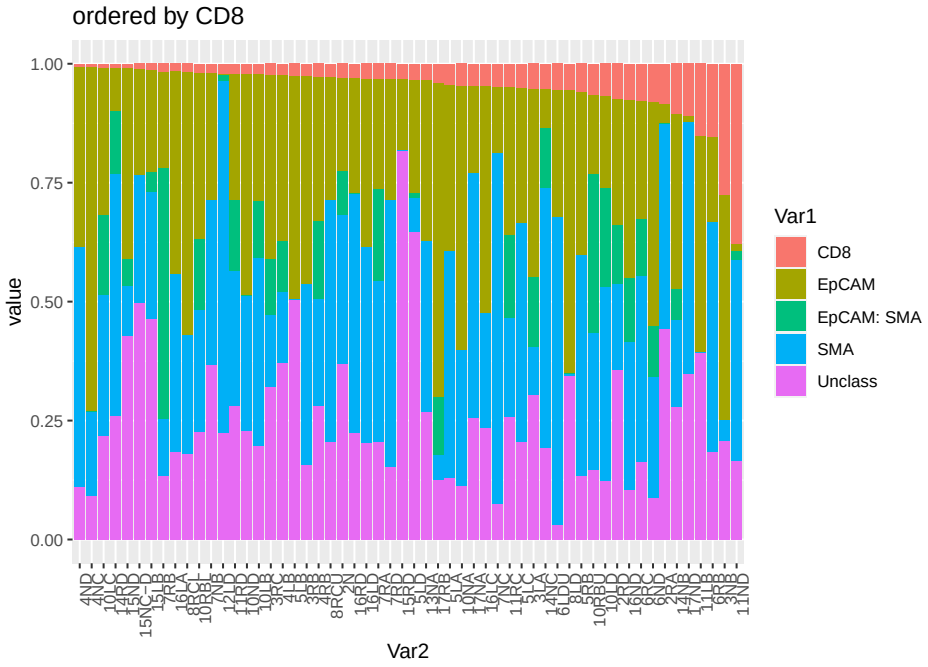
- tumor size
- growth rate
- growth rate (categorical)
- treatment
- stromal restricted or infiltrating

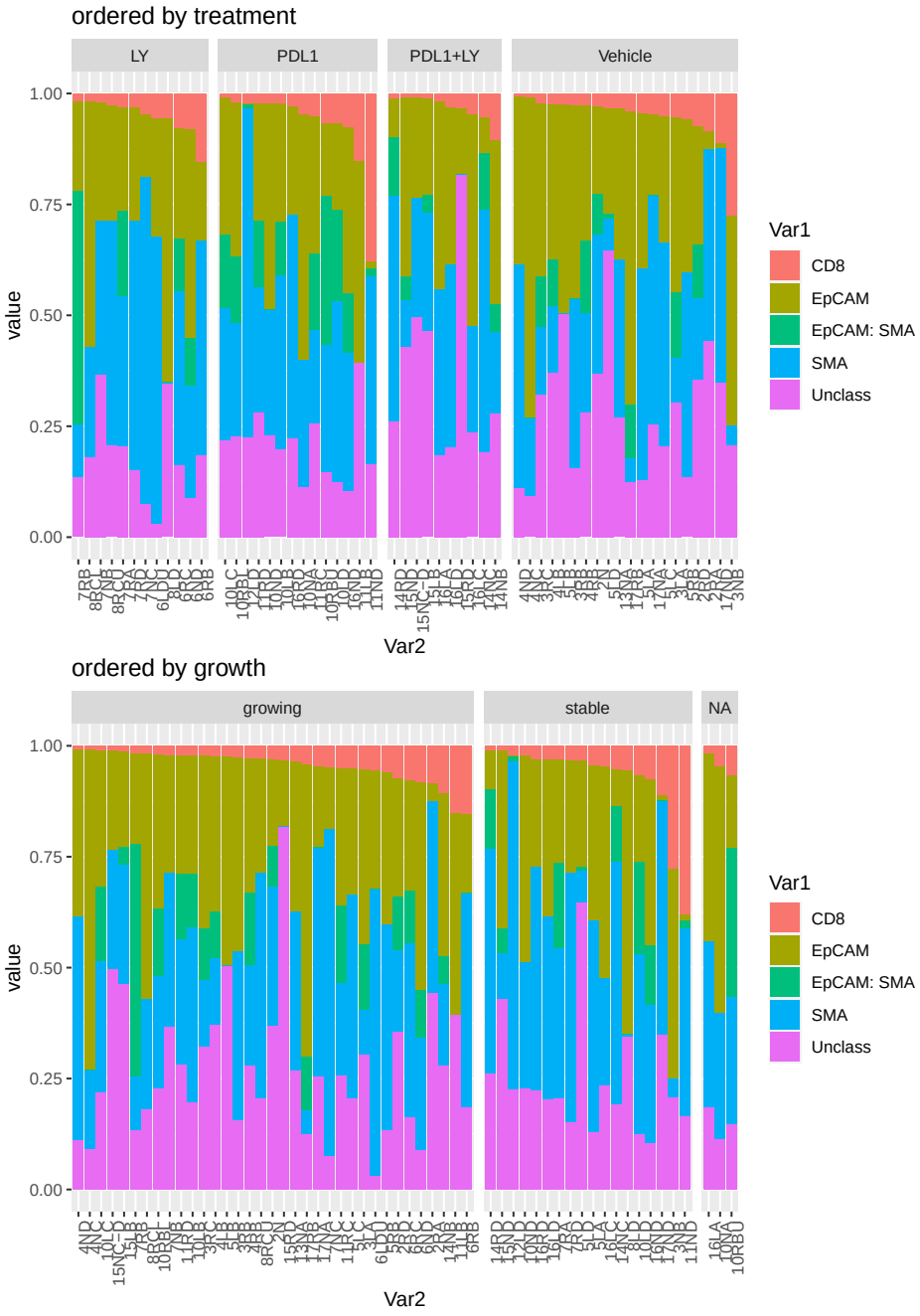
Below are the total cell counts:


ordered by CD8
value
3e+06
2e+06
1e+06
0e+00
Var2
Var1
CD8
EpCAM
EpCAM: SMA
SMA
Unclass
ordered by treatment
LY
PDL1
PDL1+LY
Vehicle
value
3e+06
2e+06
1e+06
0e+00
Var2
Var1
CD8
EpCAM
EpCAM: SMA
SMA
Unclass



Here, the same data is shown and normalised according to total cell count:

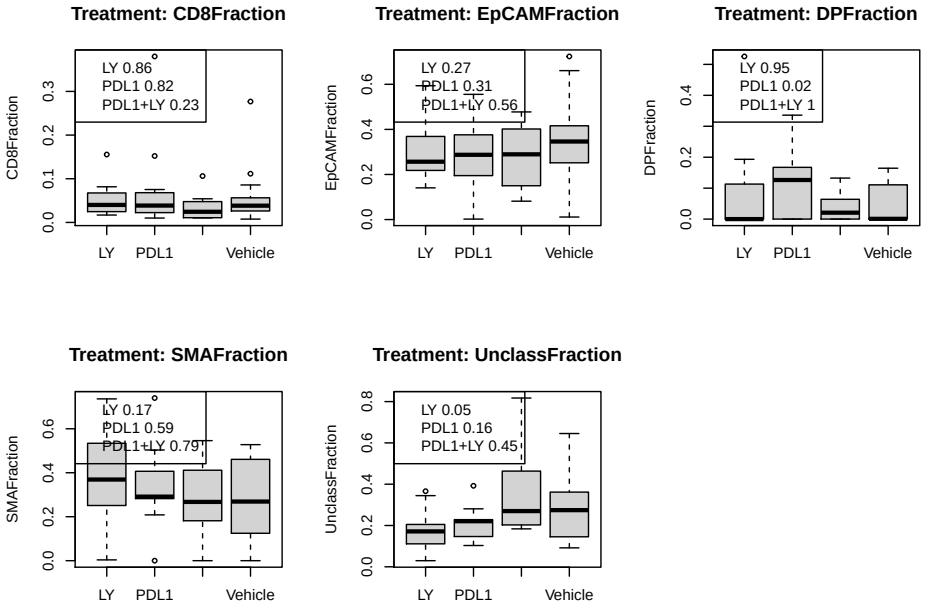



ordered by treatment
LY
PDL1
PDL1+LY
Vehicle
value
1.00
0.75
0.50
0.25
0.00
Var2
Var1
CD8
EpCAM
EpCAM: SMA
SMA
Unclass
ordered by growth
growing
stable
NA
value
1.00
0.75
0.50
0.25
0.00
Var2
Var1
CD8
EpCAM
EpCAM: SMA
SMA
Unclass

3.3 Associate composition with other covariates

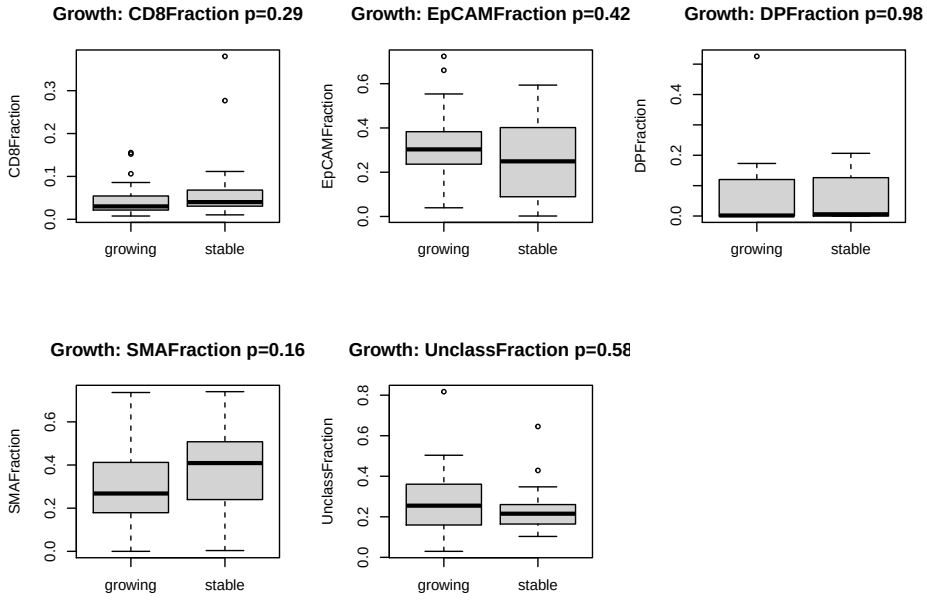
We can collapse the above data into boxplots to see if there is an association with treatments, performed using non-parametric wilcoxon test.

Compared to the vehicle, LY treated samples have a lower "stromal" fraction (unclassed DAPI+ cells) and PDL1 treated samples are more likely to have a EpCAM+SMA+ double positive fraction



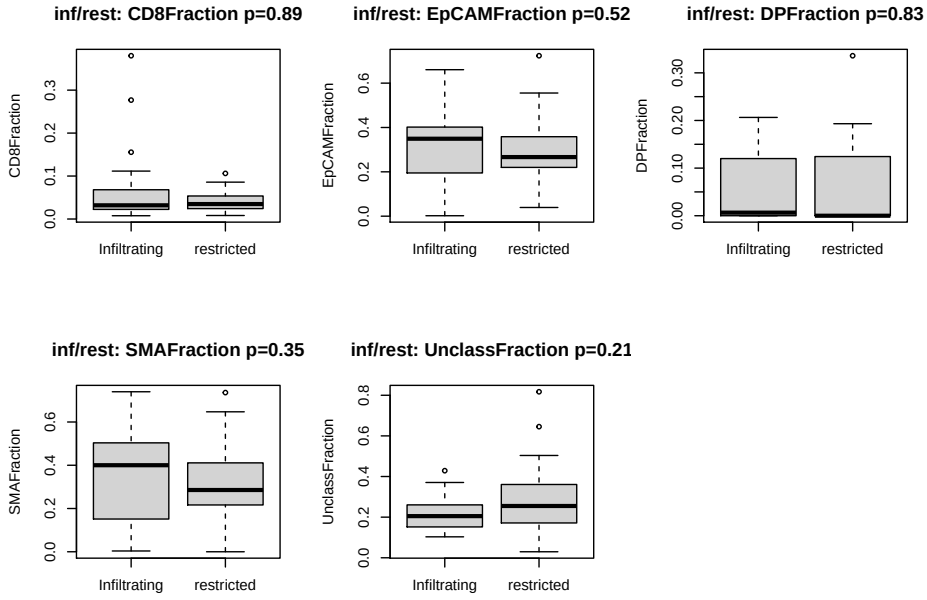
When comparing these fractions to growth, there is no association:

```
## Warning in wilcox.test.default(Cdata[Cdata$Growth2 == "growing", i],
## Cdata[Cdata$Growth2 == : cannot compute exact p-value with ties

## Warning in wilcox.test.default(Cdata[Cdata$Growth2 == "growing", i],
## Cdata[Cdata$Growth2 == : cannot compute exact p-value with ties
```



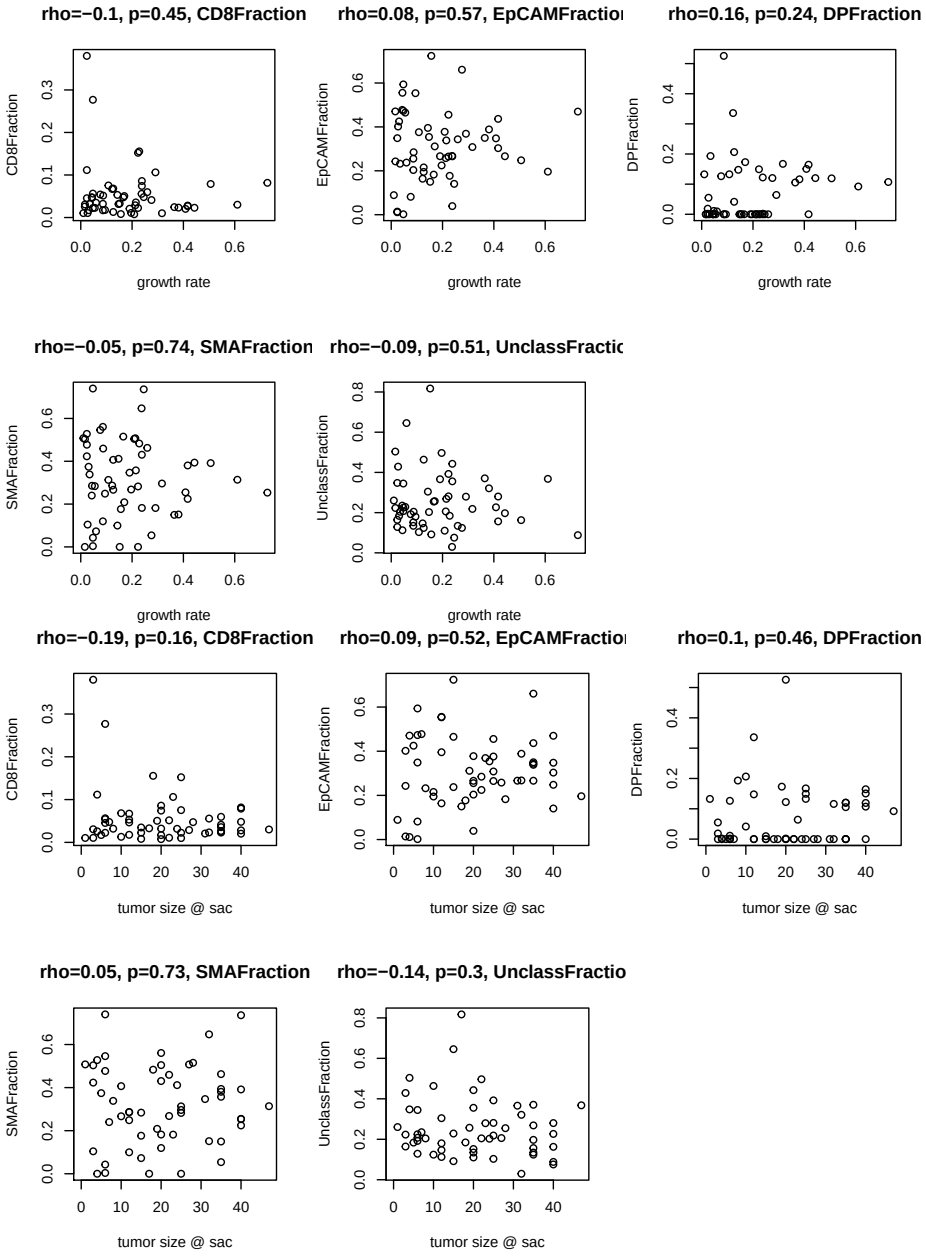
Nor is there any association with whether a sample is "immune infiltrated" or not (by manual inspection)

```
## Warning in wilcox.test.default(Cdata[Cdata$InfiltratingVsRestricted ==
## "Infiltrating", : cannot compute exact p-value with ties

## Warning in wilcox.test.default(Cdata[Cdata$InfiltratingVsRestricted ==
## "Infiltrating", : cannot compute exact p-value with ties
```



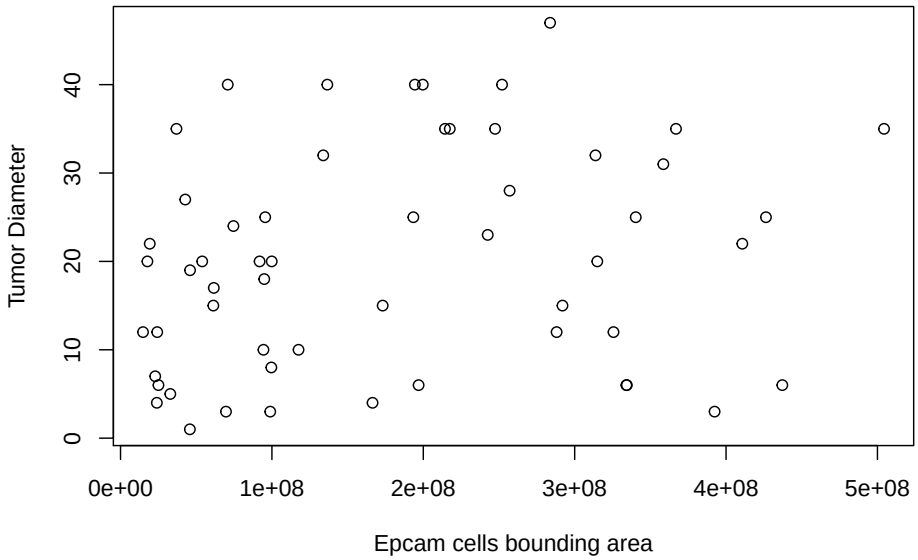
We can also use linear regression models to assess:

- correlation between growth rate and cell fractions (no associations)
- correlation between final tumor size and cell fractions (no association)



3.4 Estimate tumor size

Using WSI data, we can estimate a tumor size for each tissue sample and compare to the final tumor sizes. This will be based on the distribution of EpCAM+ cells. This estimate is also used to normalise CD8 counts earlier on.



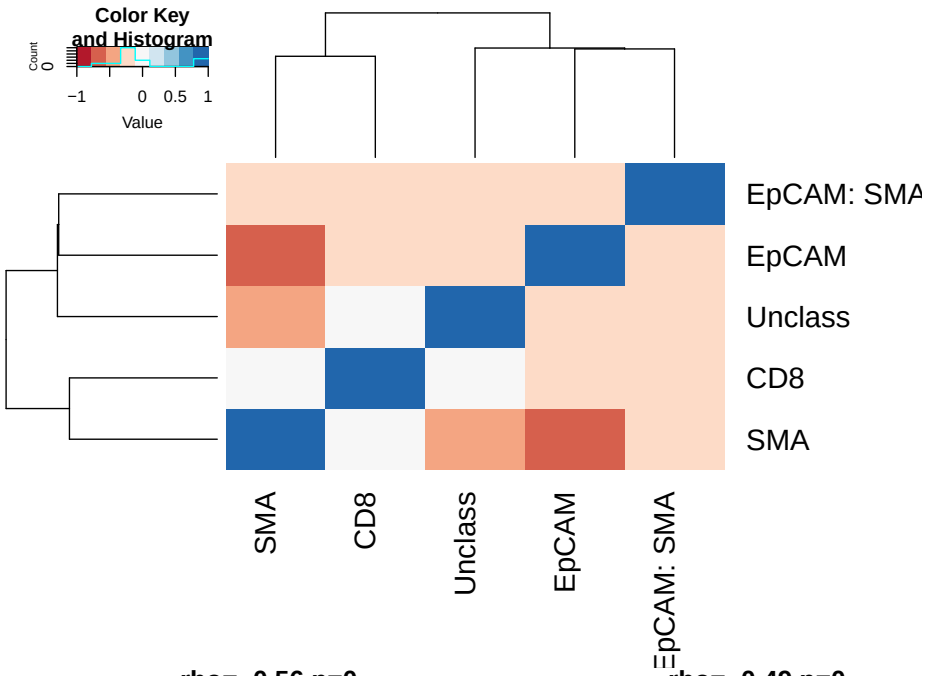

```
##
##  Pearson's product-moment correlation
##
## data:  TareaSum and Tdiameter
## t = 1.6859, df = 54, p-value = 0.09759
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04174574  0.45949702
## sample estimates:
##       cor
## 0.2236089
```

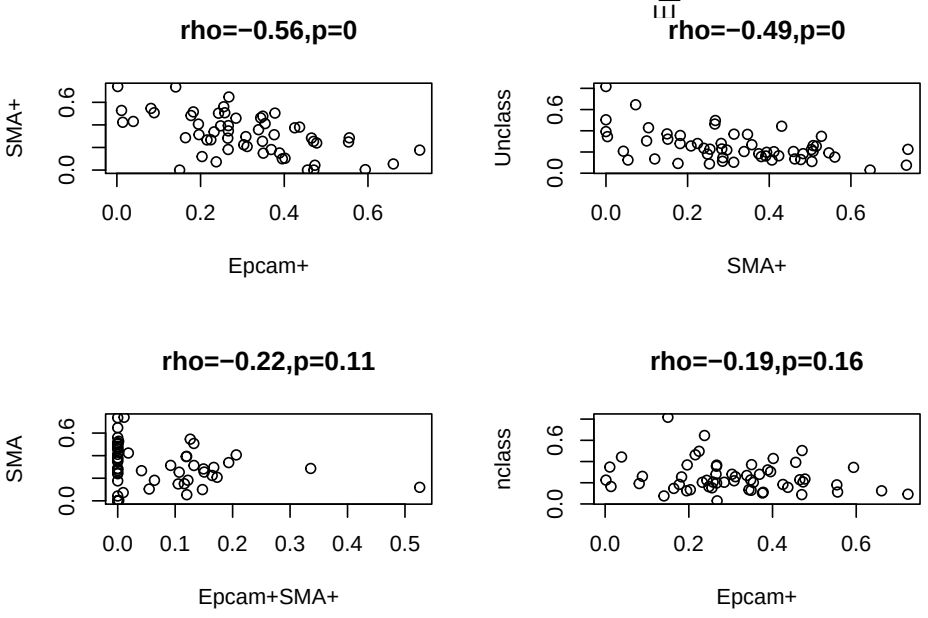
3.5 Correlations between different subpopulations

Look for correlates between different subpopulations: Naturally, we would expect a negative correlation since this should sum to 1. Below are heatmaps showing correlations between different cell types, and significant associations are linearly shown.

Note the following negative correlations:

- epcam and SMA
- SMA+ and Unclass






```
##                CD8      EpCAM EpCAM: SMA        SMA    Unclass
## CD8        1.0000000 -0.1780975 -0.1268947 -0.0344095 -0.1107879
## EpCAM     -0.1780975  1.0000000 -0.1595219 -0.5597908 -0.1898428
```

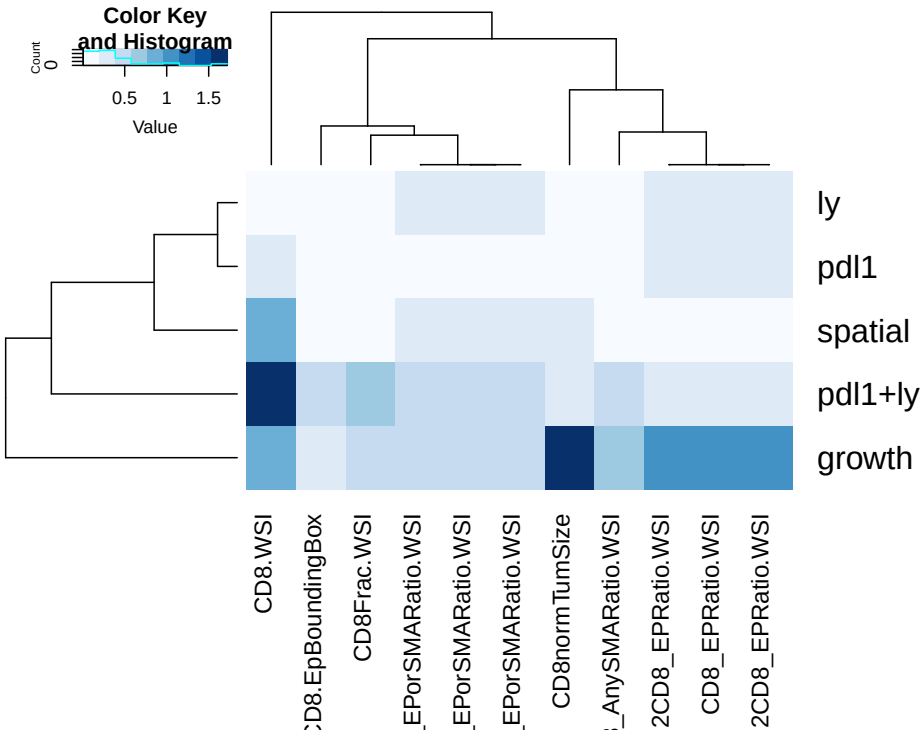
```
## EpCAM: SMA -0.1268947 -0.1595219  1.0000000 -0.2171477 -0.1783030
## SMA        -0.0344095 -0.5597908 -0.2171477  1.0000000 -0.4946850
## Unclass    -0.1107879 -0.1898428 -0.1783030 -0.4946850  1.0000000
```

```
##           [,1]         [,2]      [,3]         [,4]        [,5]
## [1,] 0.0000000 1.891064e-01 0.3513624 8.012235e-01 0.416293560
## [2,] 0.1891064 0.000000e+00 0.2402481 7.272201e-06 0.161096722
## [3,] 0.3513624 2.402481e-01 0.0000000 1.079270e-01 0.188588161
## [4,] 0.8012235 7.272201e-06 0.1079270 0.000000e+00 0.000106427
## [5,] 0.4162936 1.610967e-01 0.1885882 1.064270e-04 0.000000000
```

3.6 Associations between CD8 counts with other clinical variables

Below we assess whether any of the CD8-variables described in section 3.1 is associated with

- treatment
- growth
- spatial pattern



Note that CD8 normalised by tumor size is associated with growth (but this could a reflection of the size of the tumor), and there is a borderline differ-

ence once normalised by epithelial content. In addition the CD8 total count is associated with pdl1+ly treatment.

Note that p=0.05 is designated by a value of 1.3

Chapter 4

Spatial statistics

Below, we use three different metrics to compare spatial distributions:

- k-nearest neighbour distances
- the interacting fraction
- morisita-horn distances

These are compared to manual inspection of the result

4.1 knn-Distances:

The k-nearest neighbour distances looks at the average distance from a given cell type of class A to a cell type of class B. In this section, the reference class A is the CD8 T cell, and we will look at the mean distance to SMA, Epcam, double positive and unclassified cells in each image.

To account for potential fluctuations due to misclassified cells, or isolated single cells, k values of 1, 3, 5 will be used. I.e. for each cell, we will compute the mean distance from each Cd8Tcell to its 1, 3, and 5 nearest neighbours.

4.1.1 Comparison to manual classification, treatment, growth

Overall, we see that the differences in infiltrating vs restricted are similar. We see statistical differences (using anova followed by Tukey's test) between:

- epcam and SMA-epcam in both cases (higher distances to EpCAM on average)
- SMA-Epcam to SMA (CD8s are closer to SMA+)
- Unclass to Epcam-SMA (CD8s closer to unclass)

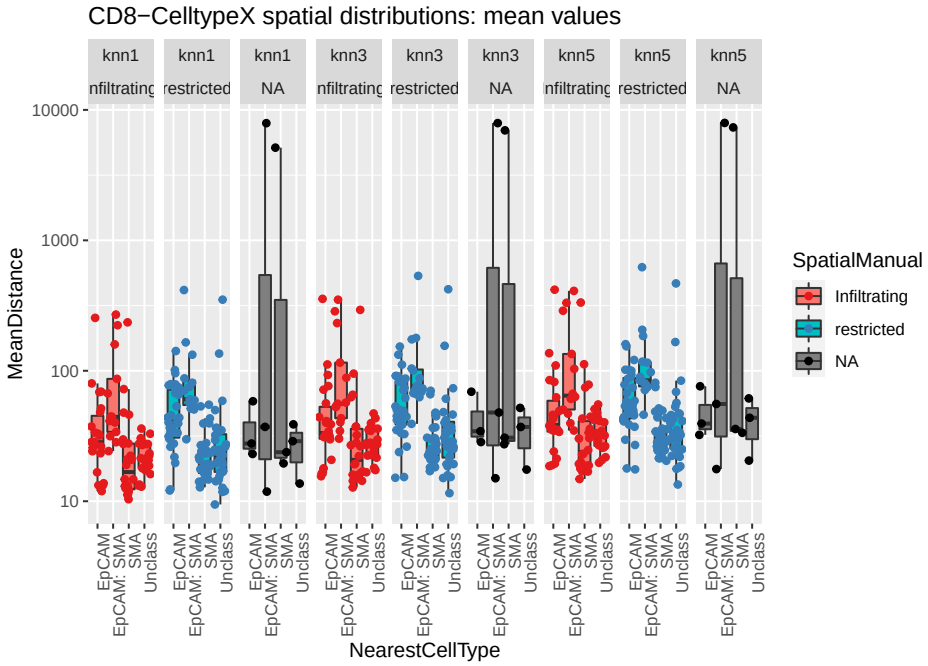
In the infiltrating case:

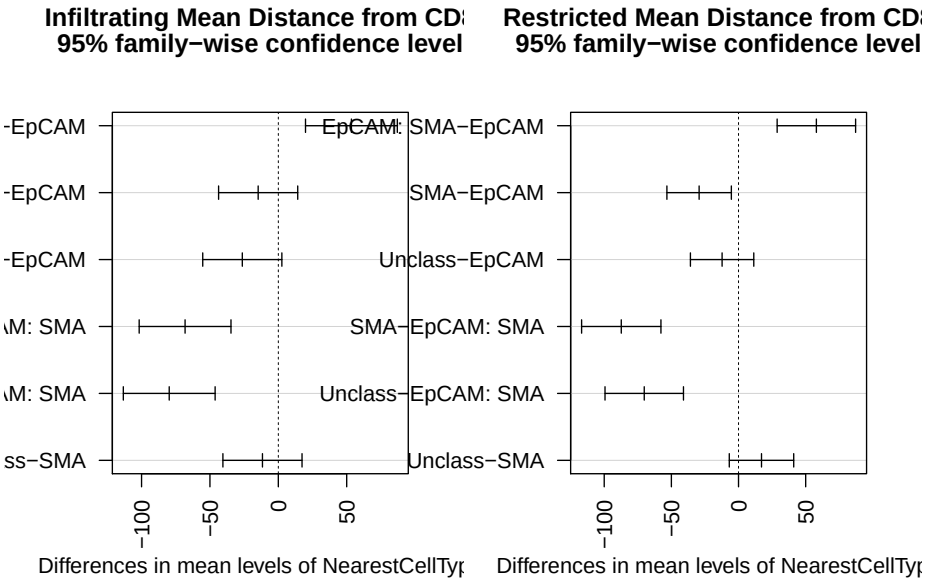
- Unclass to Epcam (CD8s closer to unclass, borderline significant)

In the restricted cases, we see:

- Unclass to SMA (higher distance to unclass in the restricted case)
- SMA to epcam (CD8s are closer to the SMA)

This last result is consistent with what we expect for a CD8+ cell which is stroma-restricted.



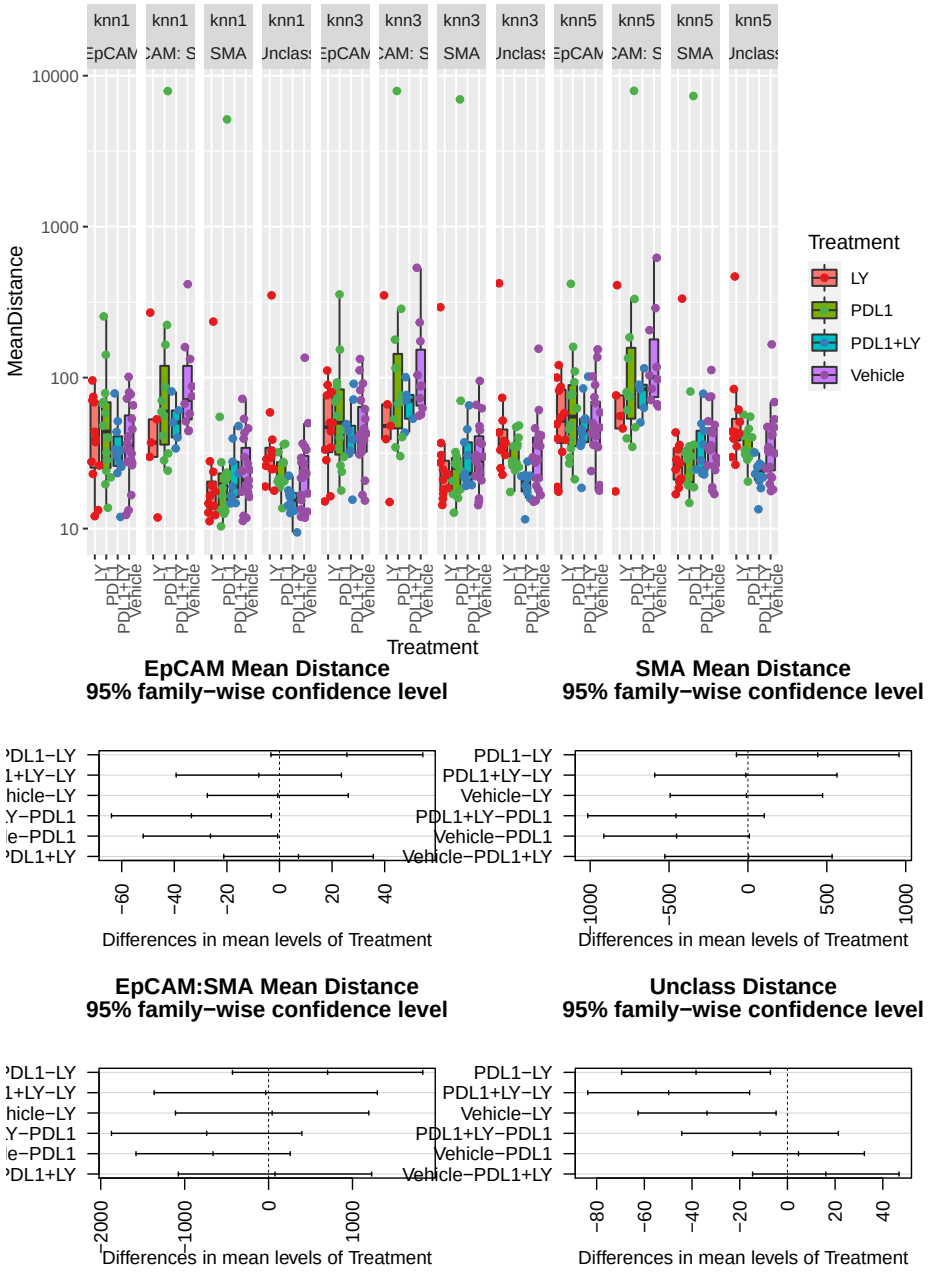


4.1.2 Associations with outcome

We can also see if there is an association between these distances with growth and treatment

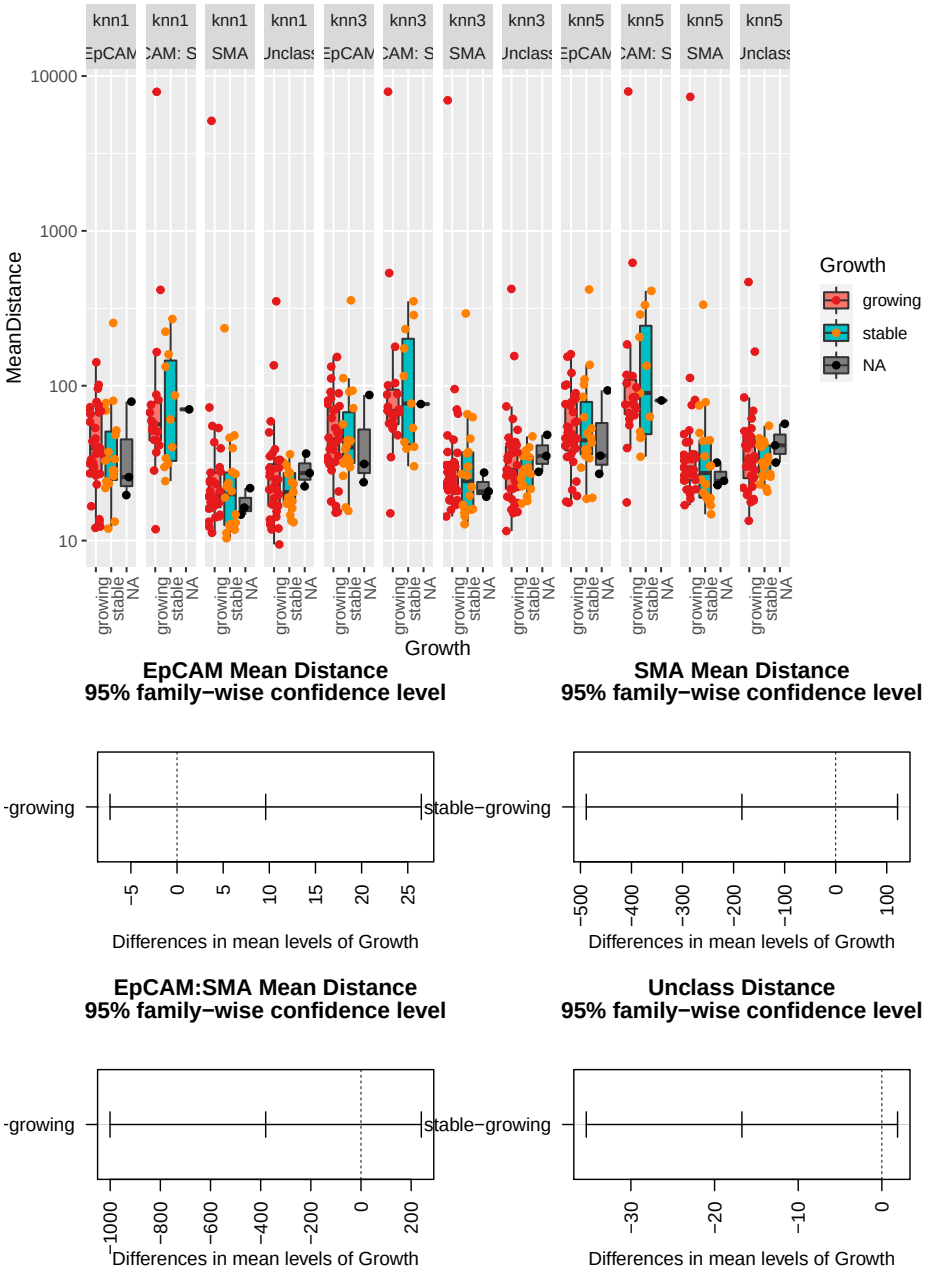
Treatment:

- CD8 cells in PDL1 sample are further away to SMA+ cells and EpCAM+ (compared to vehicle or double agent)
- CD8 cells in LY treated samples are further away from unclassified cells (compared to any of the other treatments)



4.1.3 Growth

All 95% confidence lines cross 0, but it appears that stable cases have a closer unclass-CD8 interaction distance compared to growing.

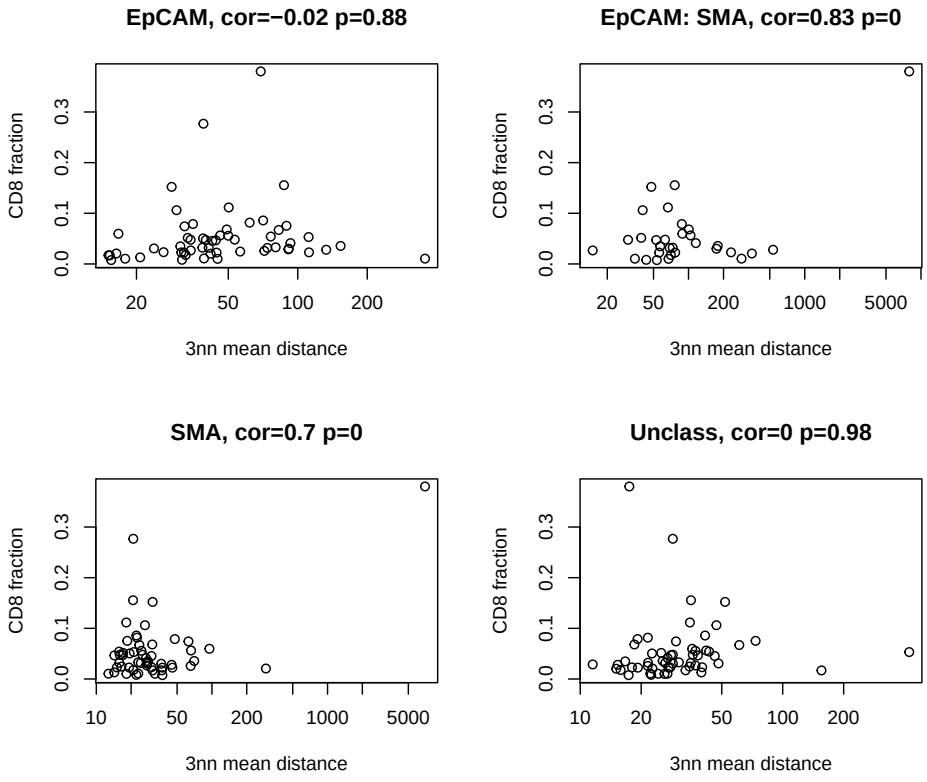


Based on the above distributions, knn1, knn3, and knn5 analyses give similar results. In the following section, we will make comparisons using knn3 results.

4.1.4 Association between distance and content

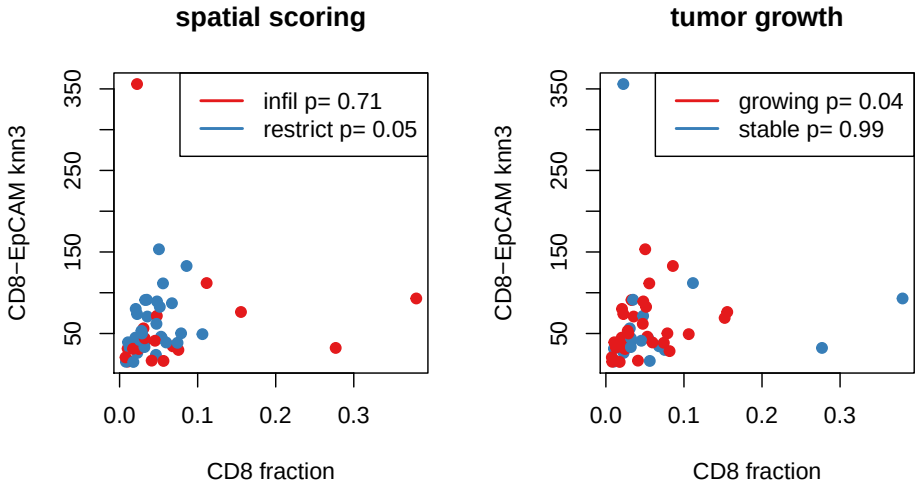
The spatial differences could be influenced by CD8 content. Here, we test if the distances from CD8Tcells to celltype B could be influenced by this.

We see that there is a correlation between CD8 fraction and the SMA proportion (both SMA and double positive). But we don't see an association with epcam+ cells or unclassified stromal cells



CD8-Epcam distances are looked at in greater detail below. Here, we look at whether there is an association with CD8 fraction in a growing and stable tumors.

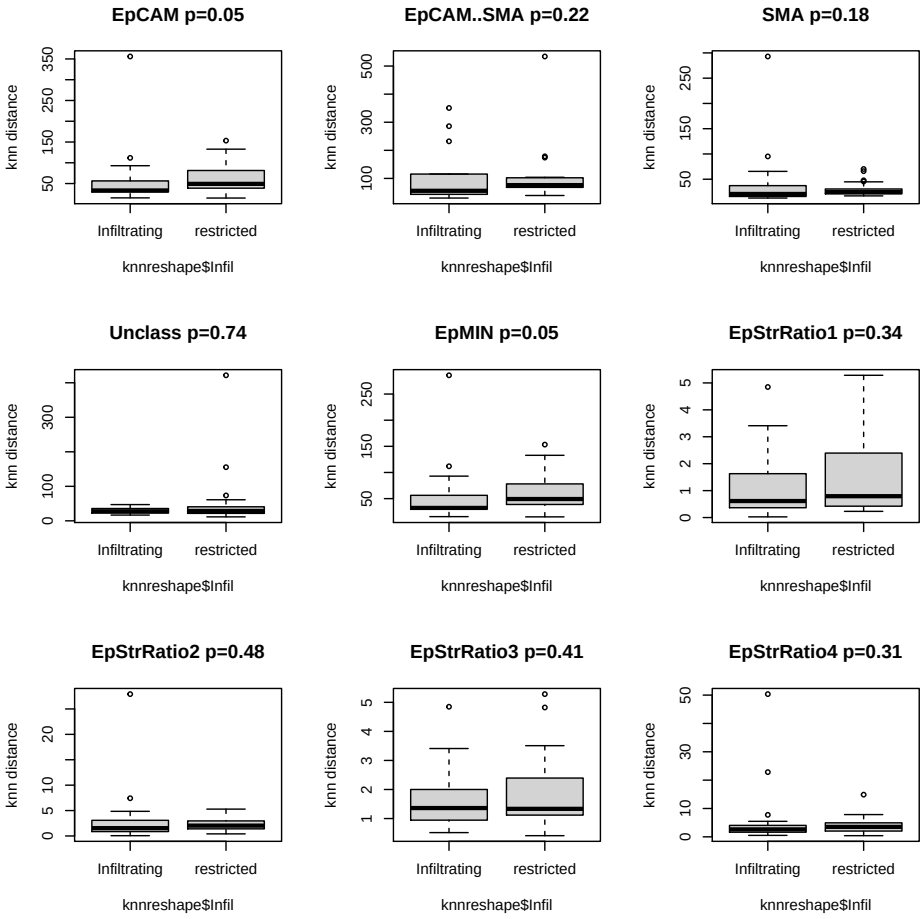
In growing or spatially restricted tumors, there is an association between the CD8-epcam distances with the CD8 fraction, but not in infiltrating or stable tumours.



Do any of the below metrics correlate with manual scoring:

- We can assess the distances from CD8 cells which come automatically:
  - Ep-distance
  - SMA distance
  - Ep:SMA distance
  - SMA distance.
- EpMIN: distance to any Epcam expressing cell (includes EP+SMA+)
- Ratios between Epcam and SMA: There are different values depending on whether the double positive fraction is used or not
  - (Ratio 1): Ep/any SMA
  - (Ratio 2): EP/SMA
  - (Ratio 3): Any Ep / Any SMA
  - (Ratio 4): Any Ep/ SMA

Whilst the manual scorings associate strongly with CD8-EPcam distances, there is no association with any other cell type.



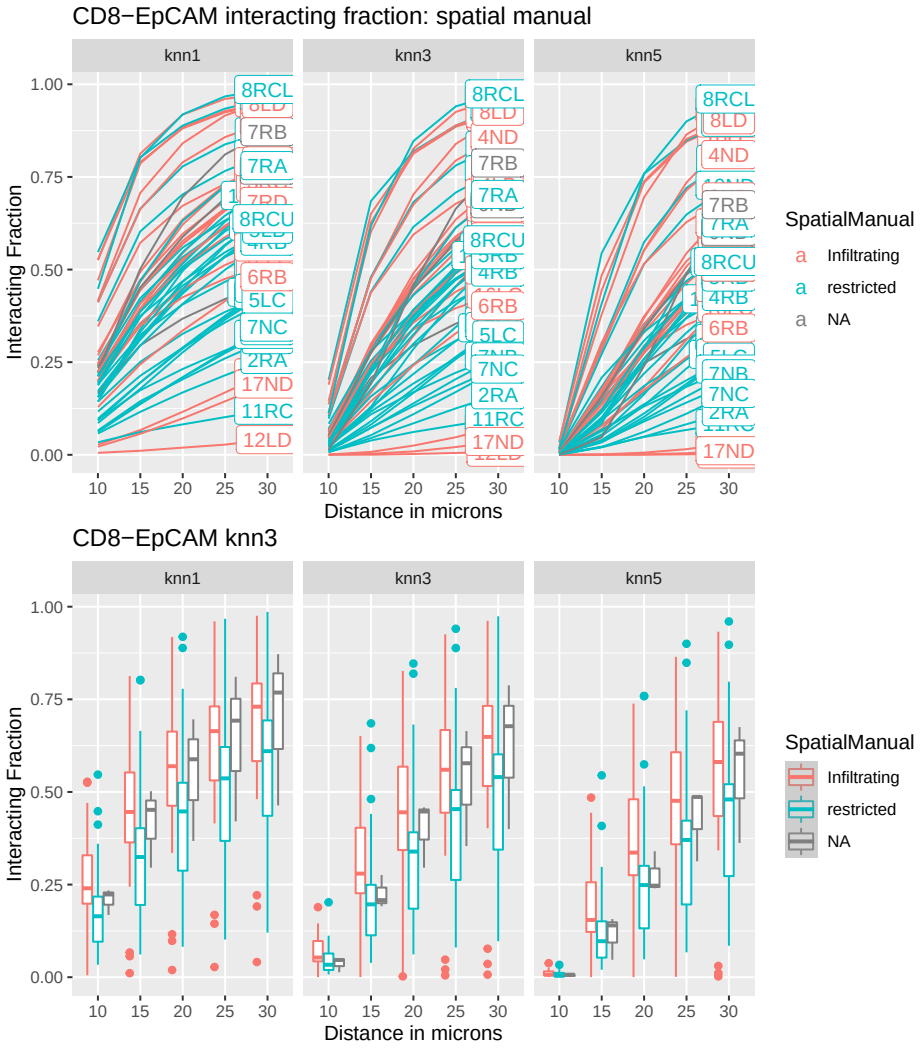
4.2 The interacting fraction

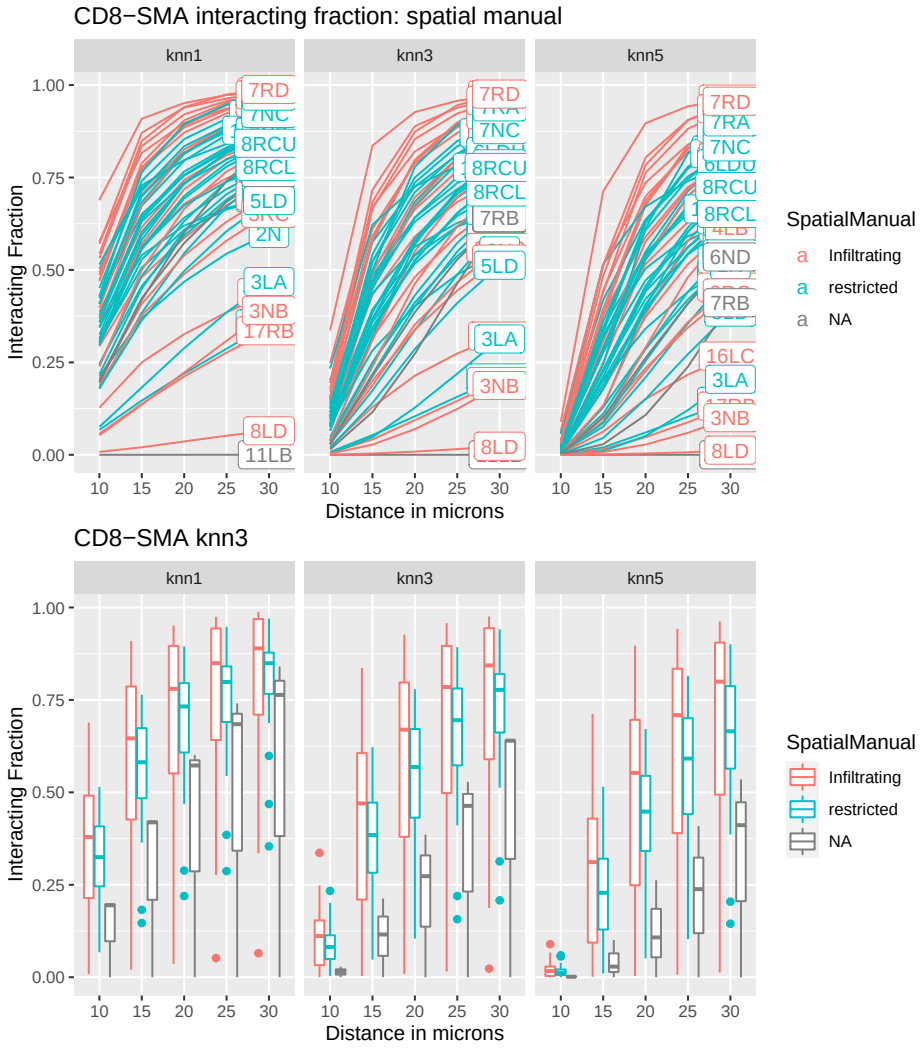
The interacting fraction uses the knn-distances and determines the proportion of CD8 cells which are within a proximity of r um from celltype B.

4.2.1 Comparison to manual & select optimal r

Below are plots of the proportion of CD8 cells within an "interacting distance" as we increase r. This looks at both the interacting fraction of CD8 cells with Epcam+ and SMA+ cells. Lines are color coded according to the manual spatial-infiltration annotation.

We notice from the line plots for each single sample that the restricted samples generally have low interacting fractins with EpCAM and SMA compared to the infiltrating samples. In addition, there is a statistical difference in EpCAM measurements compared to SMA.


CD8−EpCAM interacting fraction: spatial manual
knn1
knn3
knn5
Interacting Fraction
1.00
0.75
0.50
0.25
0.00
10
15
20
25
30
Distance in microns
8RCL
8LD
7RB
7RA
8RCU
6RB
5LC
7NC
2RA
17ND
11RC
12LD
4ND
4RB
SpatialManual
a Infiltrating
a restricted
a NA
CD8−EpCAM knn3
Infiltrating
restricted
NA



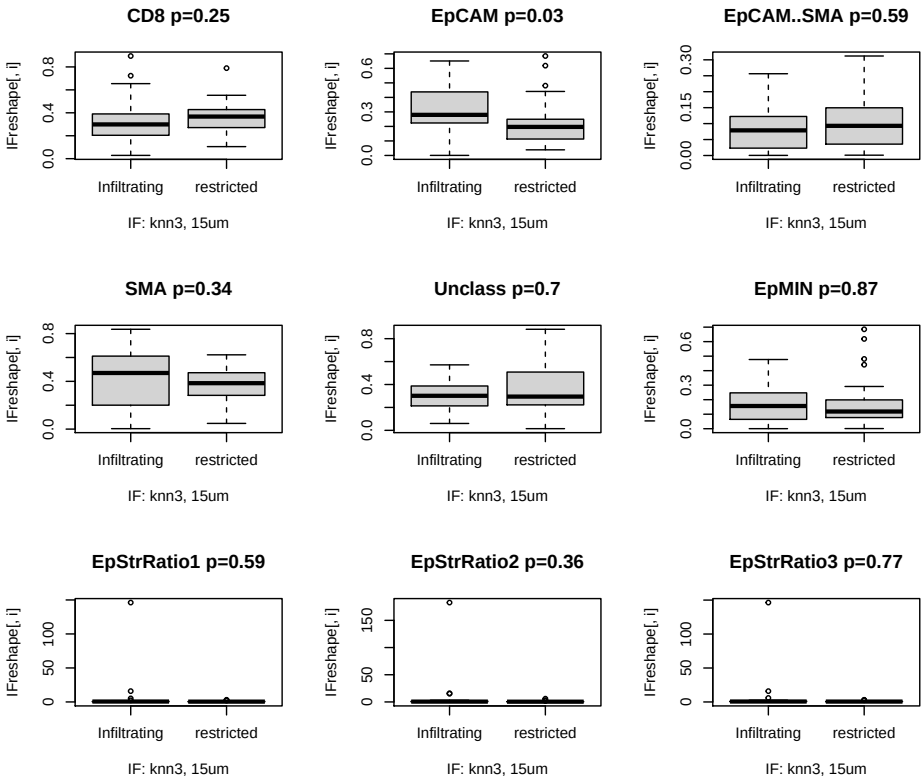
Using the boxplots as a guide, we can determine optimal "interacting distances" at which to perform downstream analysis. The best separation between restricted and infiltrating for EpCAM appears at:

- 1-nn: 10-15 um
- 3-nn: 15 um
- 5-nn: 20 um

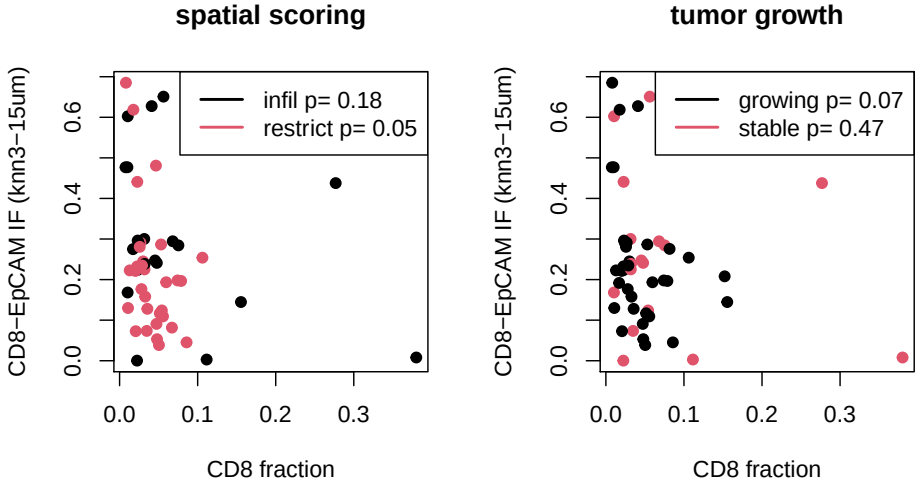
The interacting fraction does not distinguish SMA fractions (all restricted boxplots overlap with the infiltrating boxplots)

The following plots use 3NN analysis with an interacting distance of 15um. We can firstly check if there is an association between different "interacting fraction" types and manual scoring, similar to what was performed for knn-analysis. Only

CD8-EpCAM interacting distances is associated with manual scoring. All other metrics are not significant.

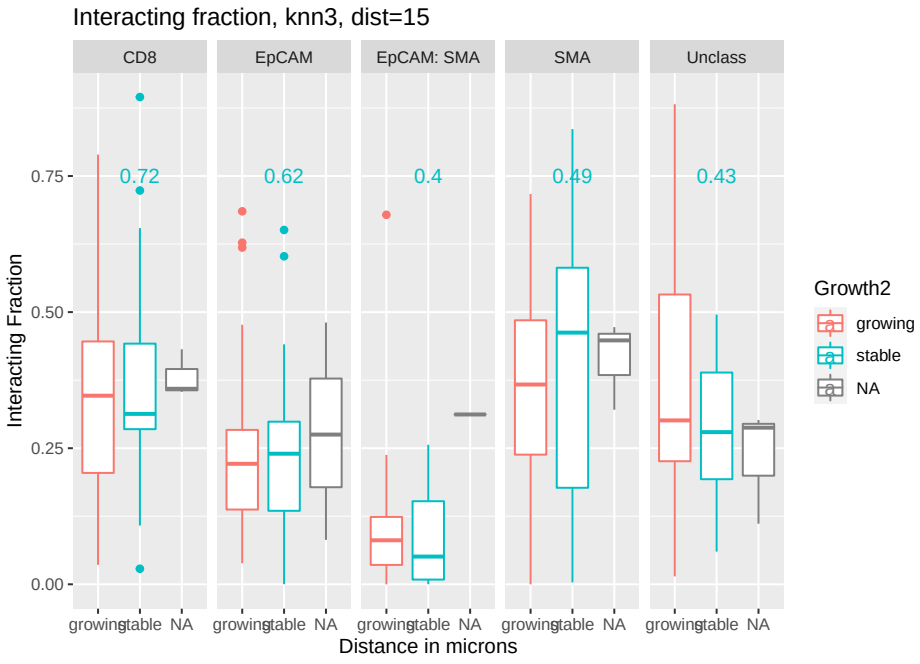


Again, association between CD8 content and interacting fraction was observed ONLY in the growing samples or restricted cases.



4.2.2 Growth

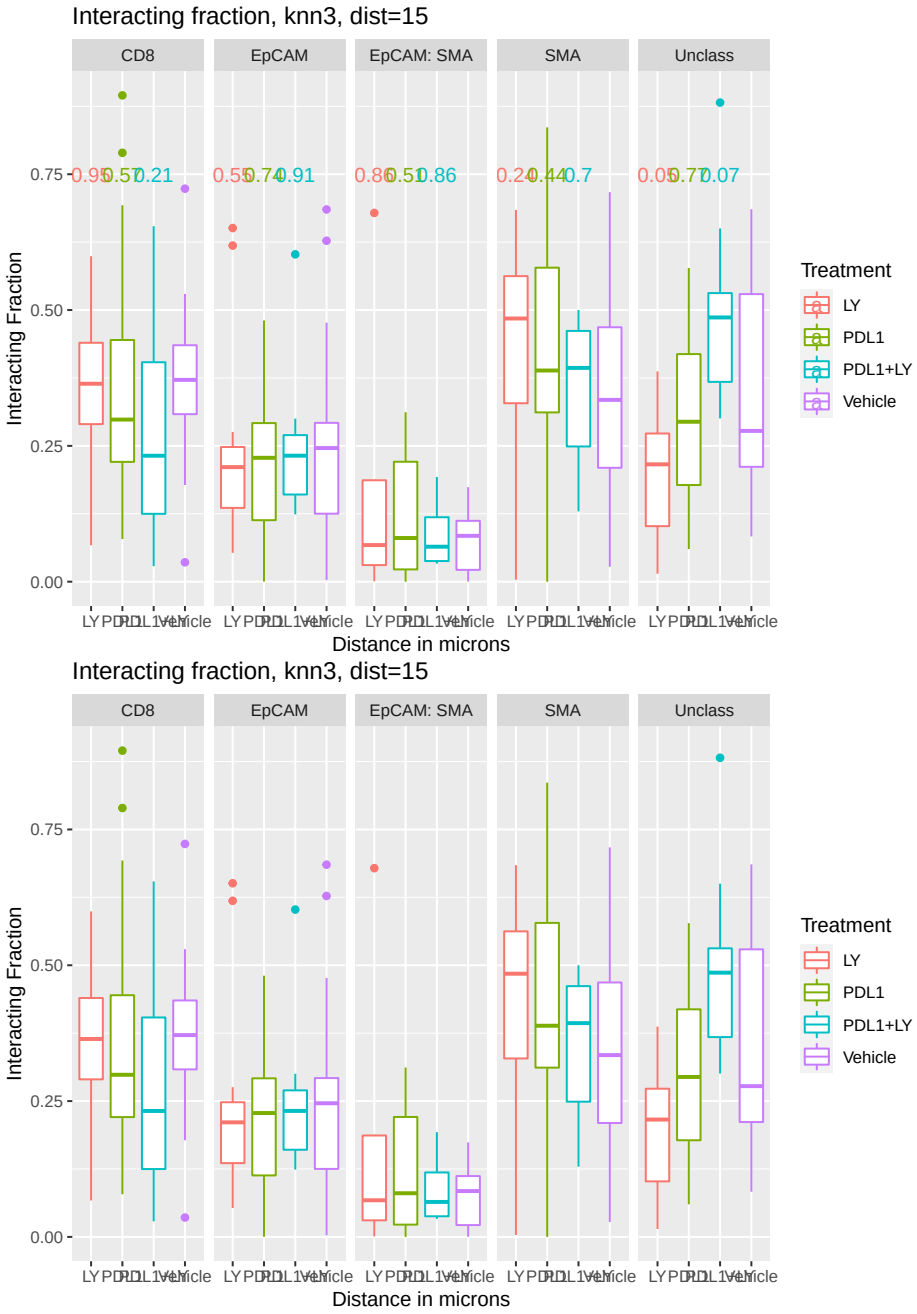
Here, we check if there is an association between the spatial pattern and tumor growth.



None of the above metrics associate with growth. (P value by wilcox test shown)

4.2.3 Treatment

Similarly, compare the distances with treatment:



There appears to be a difference in CD8-unclass interactions in LY treated samples (LY, PDL1+LY), but not in the other cases

4.3 M-H distances

The M-H distance (or Morisita Horn index) can be considered as a correlation coefficient in spatial distribution between cell type A and cell type B. To calculate this metric, the whole slide image is divided into grids of size 50 to 500um. Within each grid, the total number of cells A and B are determined.

The M-H index is thus determined as:

$$\frac{2\sum_{i=1}^{n} a_i b_i}{(D_a + D_b)AB}$$

where a_i and b_i are the number of cells in grid i,

$$A$$

and

$$B$$

the total number of cells, and

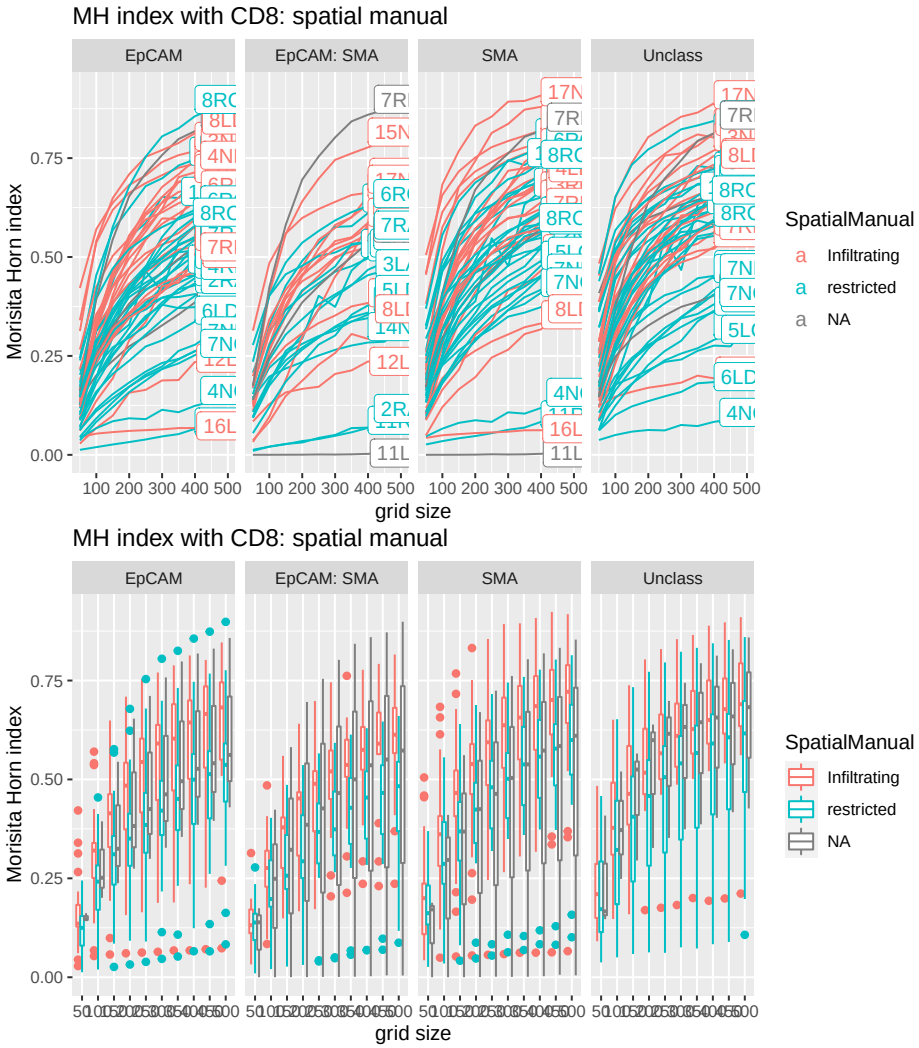
$$D_x$$

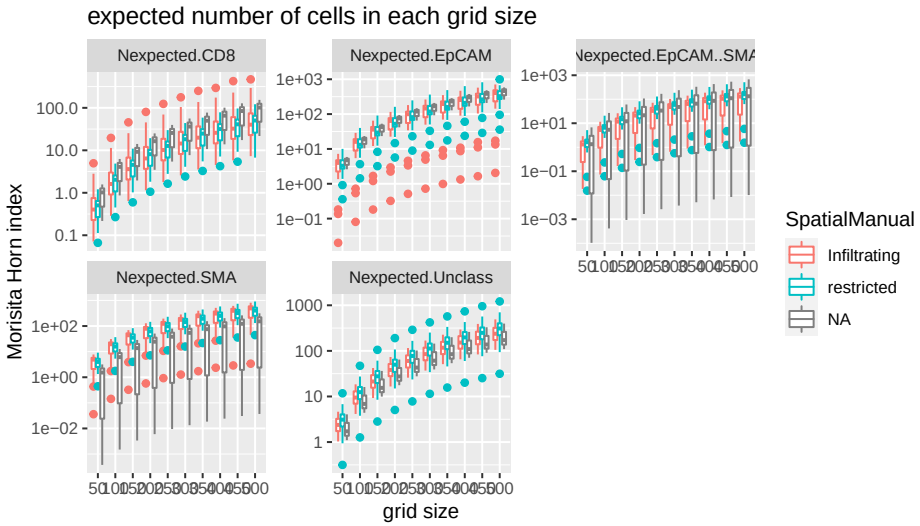
is the Simpson's index.

4.3.1 Comparison to Manual Scoring

Similar to the interacting fraction, we plot the MH index for increasing values of gridsize to determine an optimal metric to compare spatial patterns. Ideally, we would pick a metric has the following properties:

- good separation of the different values
- a reasonable number of cells within each grid (avoid too small grids which give counts of 0)
- avoid plateauing of MH values because the grid size is too large


MH index with CD8: spatial manual
EpCAM
EpCAM: SMA
SMA
Unclass
Morisita Horn index
0.00
0.25
0.50
0.75
100 200 300 400 500
grid size
SpatialManual
a Infiltrating
a restricted
a NA
MH index with CD8: spatial manual
EpCAM
EpCAM: SMA
SMA
Unclass
Morisita Horn index
0.00
0.25
0.50
0.75
grid size
SpatialManual
Infiltrating
restricted
NA



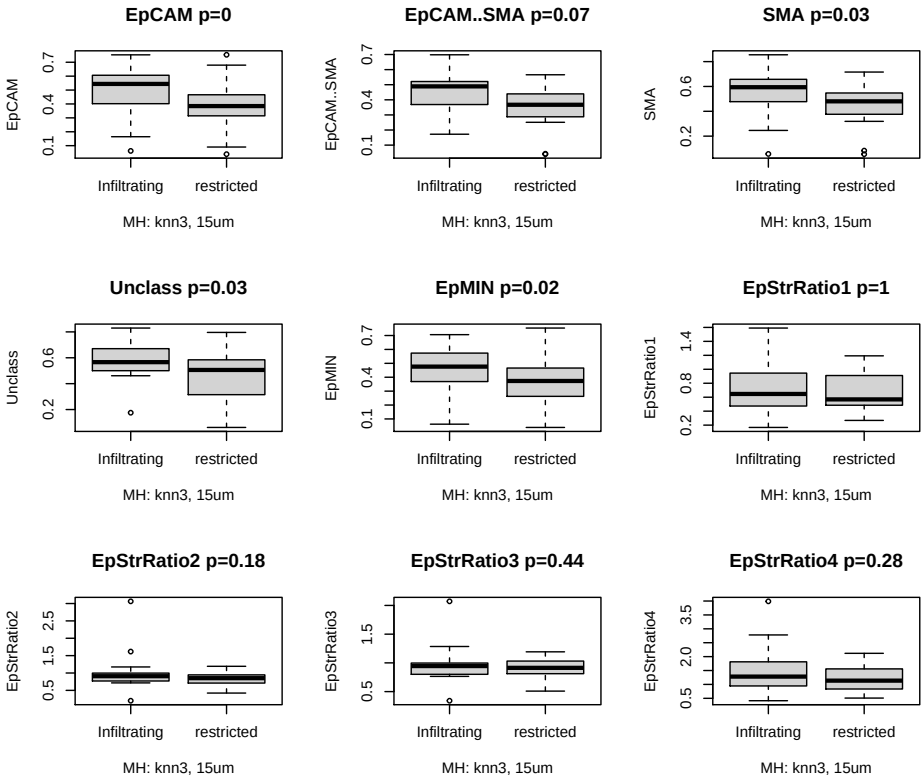
With increasing grid size, optimal differences between infiltrating and restricted appear at the following sizes:

- epCAML 100um+
- epcam:SMA most significant at 350+
- SMA: 150+
- Unclass:200+

We probably want to use a metric/gridsize of 250 um as the expected/mean number of cells in each grid is 10 here. Other notes:

- double positive cells (EpCAM+SMA+) appear in predominantly the restricted cases?
- higher SMA- stromal cells in restricted

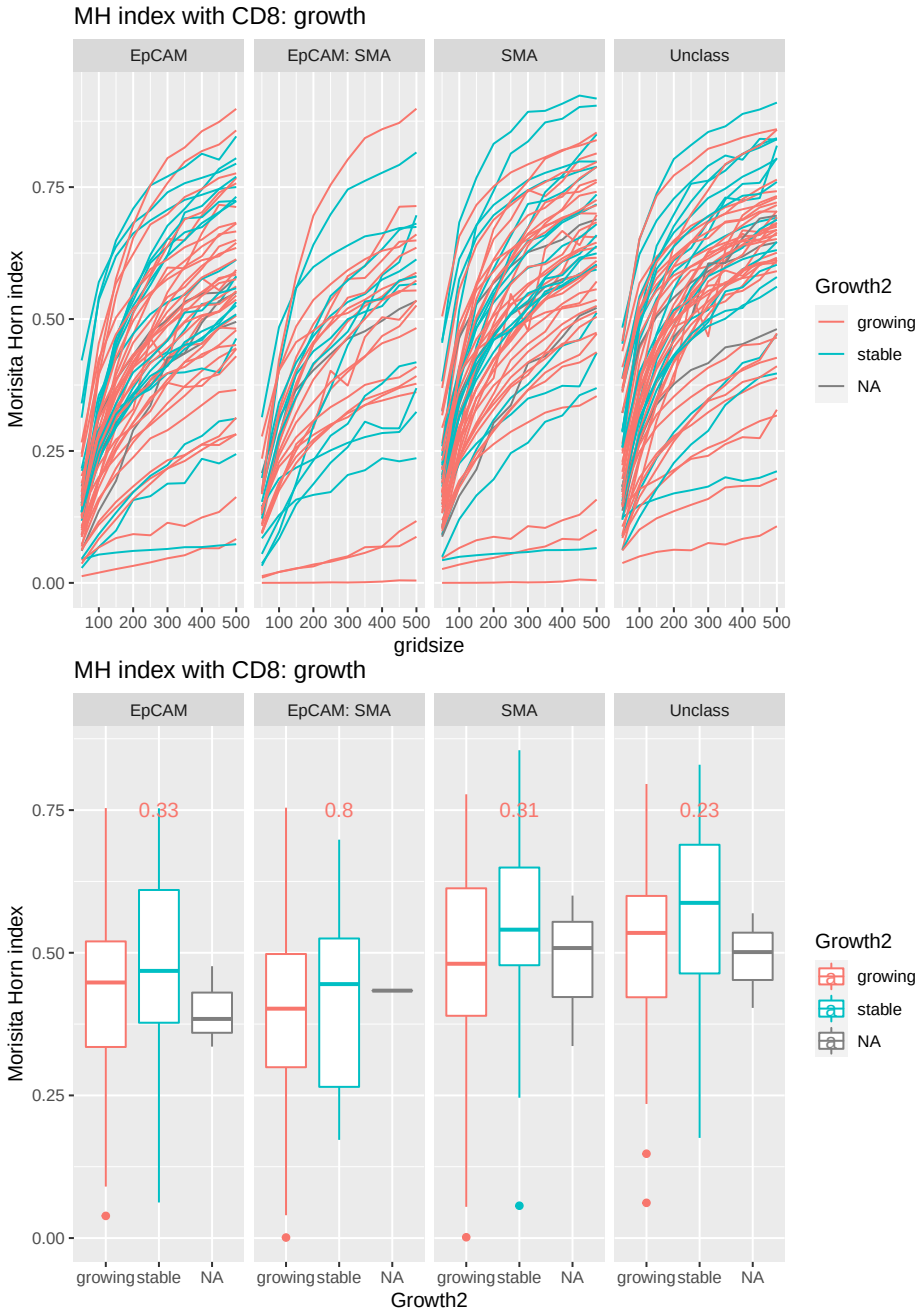
Below, we check whether these metrics can differentiate between infiltrating and restricted tumors:



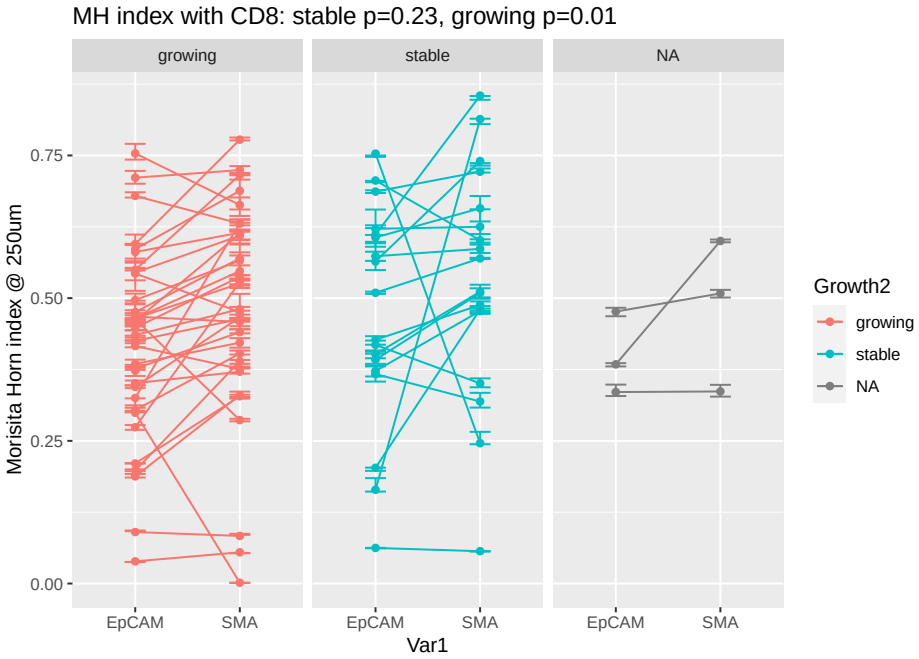
The above analysis shows that the MH-index shows differences in mixing between CD8 cells and mixing with most cell types. However MH-CD8-to-Epcam to MH-CD8-to-stroma ratios do not support this difference.

4.3.2 Growth

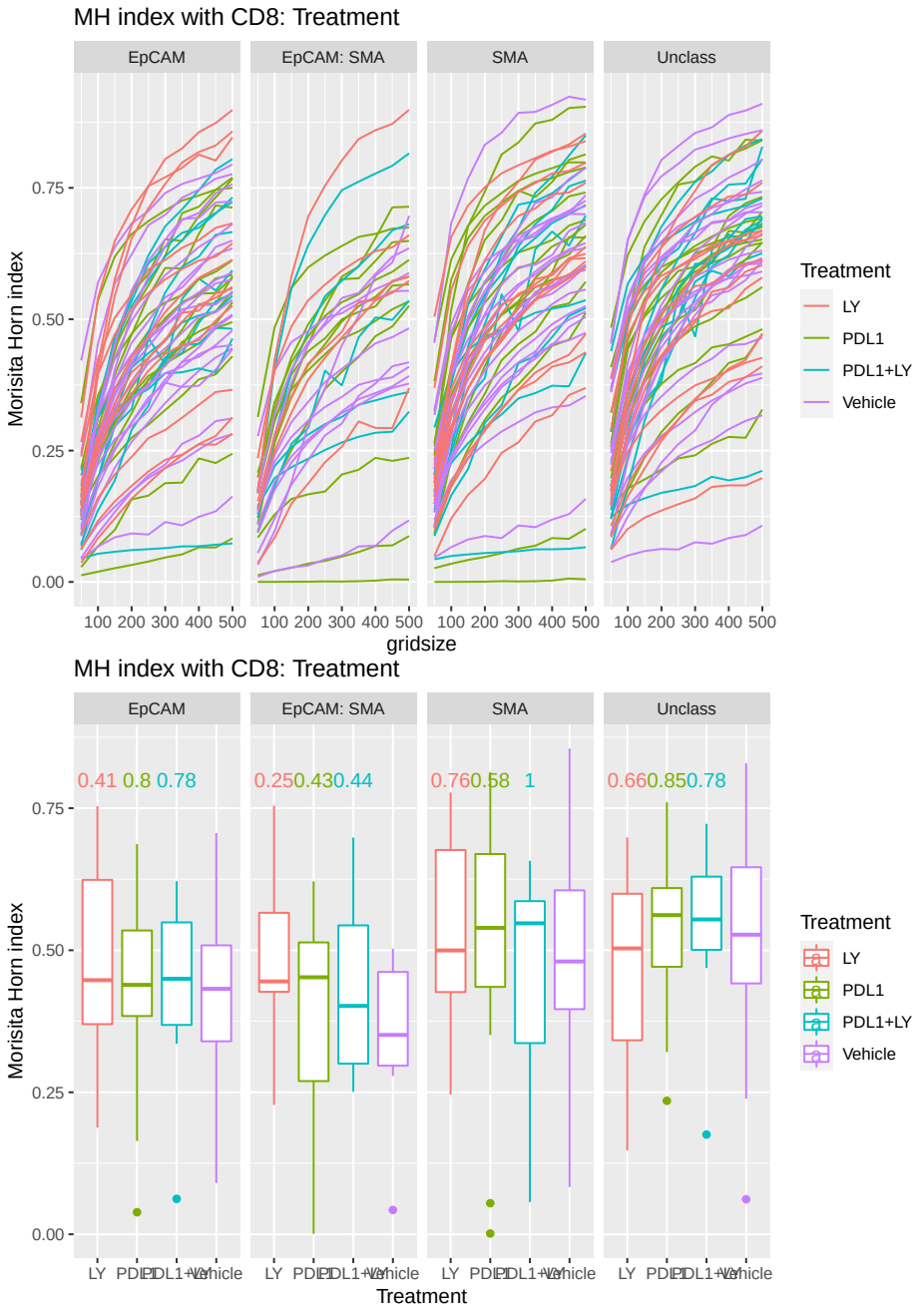
Below are the MH indices with increasing grid-size for individual samples. In general, there is a subset of stable samples which have very high intermixing



Although the MH values for growing vs stable are not different, we can compare the mixing in epcam vs stroma in matched samples:

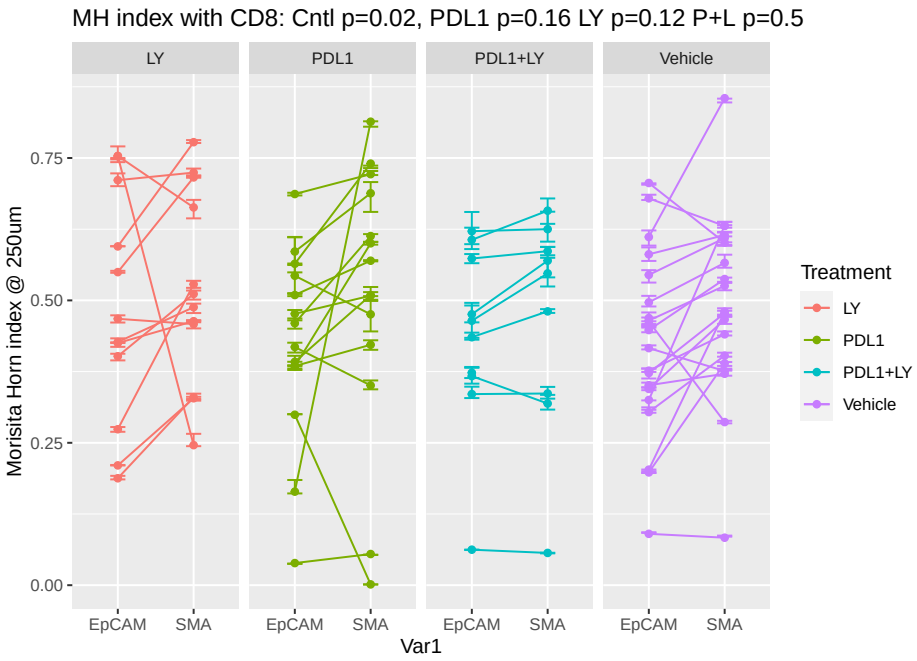

MH index with CD8: stable p=0.23, growing p=0.01
growing
stable
NA
Morisita Horn index @ 250um
0.75
0.50
0.25
0.00
EpCAM
SMA
Var1
Growth2
growing
stable
NA

4.3.3 Treatment



There is no difference between the different spatial metrics compared to the vehicle, however, we can compare for a given treatment if there is a difference between the epcam and the stromal interaction scores. It appears that there

is a difference only in the control, where SMA mixing is higher than EPcam mixing:



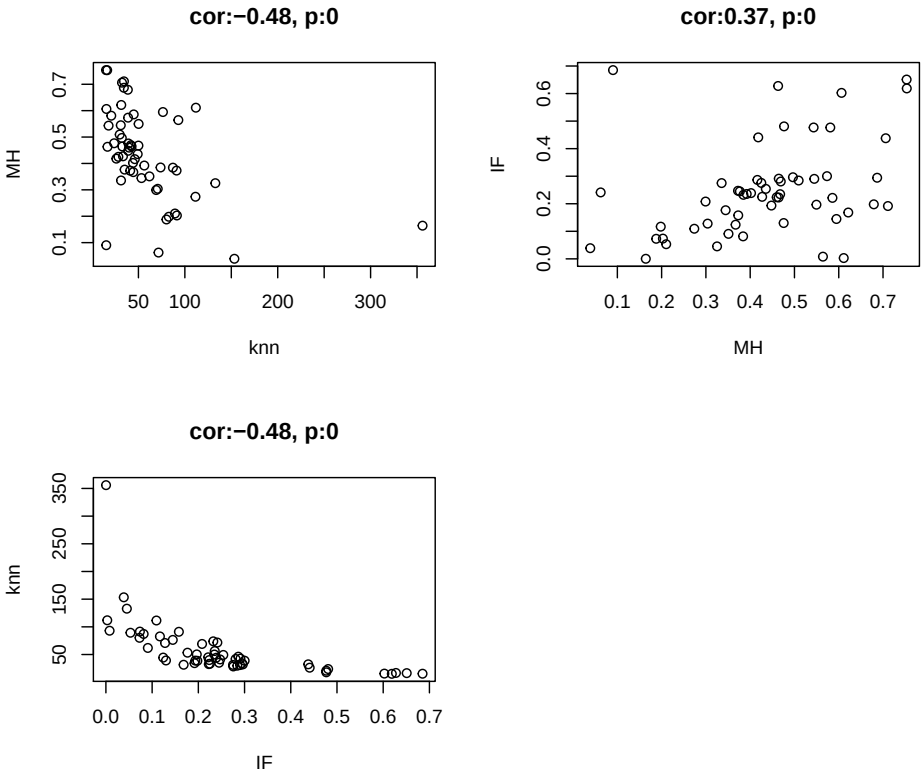
4.4 Comparison between metrics

After assessing optimal parameters for each metric, in this section we assess which metric could be the best for spatial analysis.

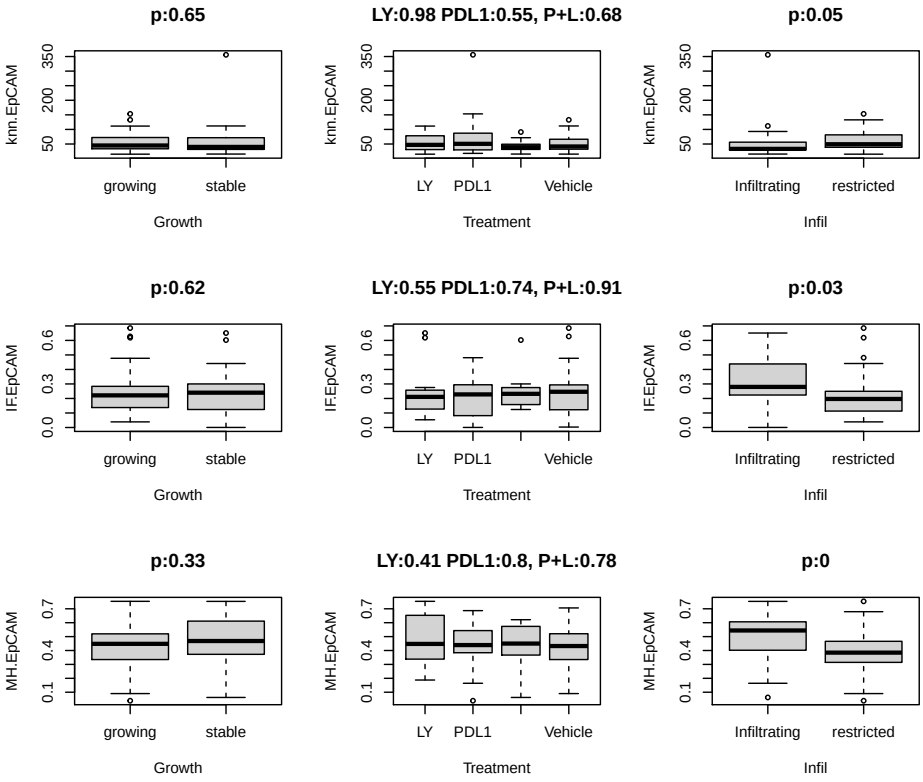
Below is a table of the different metrics and their values for each sample

	knn.EpCAM	IF.EpCAM	MH.EpCAM	Treatment	Growth	Infil	CD8fra
10LB	39.96193	0.2236555	0.4598790	PDL1	growing	Infiltrating	0.022897
10LC	17.87345	0.4768745	0.5432610	PDL1	growing	Infiltrating	0.010034
10LD	34.33110	0.2944373	0.6867357	PDL1	stable	Infiltrating	0.068108
10NA	23.80977	0.4808547	0.4764461	PDL1	NA	restricted	0.046562
10ND	26.20801	0.4408878	0.4181843	PDL1	stable	restricted	0.022719
10RBL	44.92703	0.2213380	0.5857603	PDL1	growing	restricted	0.020271
10RBU	87.12901	0.0816202	0.3840030	PDL1	NA	restricted	0.067064
11LB	69.10296	0.2080105	0.2992537	PDL1	growing	NA	0.152177
11ND	92.98237	0.0079816	0.5643066	PDL1	stable	Infiltrating	0.380067
11RC	153.33513	0.0386688	0.0388372	PDL1	growing	restricted	0.050418
11RD	73.82461	0.2323507	0.3847338	PDL1	growing	restricted	0.022445
12LD	355.94915	0.0002809	0.1642608	PDL1	stable	Infiltrating	0.022369
13NA	70.80101	0.1279338	0.3039013	Vehicle	growing	restricted	0.035581
14NB	49.22243	0.2540062	0.4353436	PDL1+LY	growing	restricted	0.106152
14NC	44.46686	0.1239597	0.3670808	PDL1+LY	stable	restricted	0.054129
14RD	31.50346	0.1681860	0.6214897	PDL1+LY	stable	Infiltrating	0.010337
15LB	32.69579	0.2226772	0.4642353	PDL1+LY	growing	restricted	0.012961
15NC-D	39.22818	0.1299775	0.4756601	PDL1+LY	growing	restricted	0.010787
15ND	15.54429	0.6025030	0.6063571	PDL1+LY	stable	Infiltrating	0.010472
15RD	91.07517	0.1578969	0.3729870	PDL1+LY	growing	restricted	0.032831
16LA	31.26867	0.2750368	0.3355138	PDL1+LY	NA	Infiltrating	0.017152
16LC	71.64542	0.2411490	0.0623470	PDL1+LY	stable	Infiltrating	0.047692
16LD	38.85862	0.3001624	0.5734648	PDL1+LY	stable	Infiltrating	0.031616
16ND	29.84045	0.2842939	0.5094333	PDL1	stable	Infiltrating	0.075330
16RD	56.31604	0.2354937	0.3920871	PDL1	stable	Infiltrating	0.030648
17NA	61.86642	0.0907525	0.3510232	Vehicle	growing	restricted	0.047268
17ND	111.85371	0.0029328	0.6112631	Vehicle	stable	Infiltrating	0.111526
17RB	16.70828	0.6275561	0.4630693	Vehicle	growing	Infiltrating	0.041107
2N	35.15413	0.2448721	0.3766073	Vehicle	growing	restricted	0.030104
2RA	132.80795	0.0452766	0.3252085	Vehicle	growing	restricted	0.085819
2RD	38.61221	0.1979009	0.6791816	Vehicle	growing	restricted	0.074219
3LA	46.09601	0.2865240	0.4162200	Vehicle	growing	restricted	0.053124
3NB	32.31159	0.4377911	0.7061056	Vehicle	stable	Infiltrating	0.276829
3RB	42.63162	0.2910751	0.4645950	Vehicle	growing	restricted	0.026655
3RC	32.11938	0.2962108	0.4965437	Vehicle	growing	Infiltrating	0.023213
4LB	30.99117	0.2903294	0.5447258	Vehicle	growing	Infiltrating	0.024447
4NC	15.33106	0.6851436	0.0901870	Vehicle	growing	restricted	0.008138
4ND	20.75556	0.4768080	0.5809047	Vehicle	growing	Infiltrating	0.007624
4RB	53.31478	0.1765476	0.3445324	Vehicle	growing	restricted	0.028009
5LA	41.12871	0.2471295	0.3724930	Vehicle	stable	Infiltrating	0.045395
5LB	41.91177	0.2807178	0.4690510	Vehicle	growing	restricted	0.025783
5LC	82.68448	0.1168545	0.1980203	Vehicle	growing	restricted	0.051505
5LD	91.41539	0.0735101	0.2028472	Vehicle	stable	restricted	0.034810
5RB	39.00602	0.1934315	0.4479302	Vehicle	growing	restricted	0.059667
6LDU	111.44785	0.1092889	0.2738092	LY	growing	restricted	0.055625
6ND	28.41556	0.2756628	0.4251078	LY	growing	NA	0.081590
6RB	76.37955	0.1446840	0.5947703	LY	growing	Infiltrating	0.155480
6RC	50.21453	0.1966022	0.5496400	LY	growing	restricted	0.078801
7NB	80.17580	0.0728176	0.1877309	LY	growing	restricted	0.020673
7NC	89.34399	0.0531233	0.2103338	LY	growing	restricted	0.047945
7RA	33.35958	0.2251296	0.4269972	LY	stable	restricted	0.031931
7RB	34.39737	0.1917334	0.7110689	LY	growing	NA	0.016946
7RD	44.13770	0.2386457	0.4019270	LY	stable	Infiltrating	0.032169

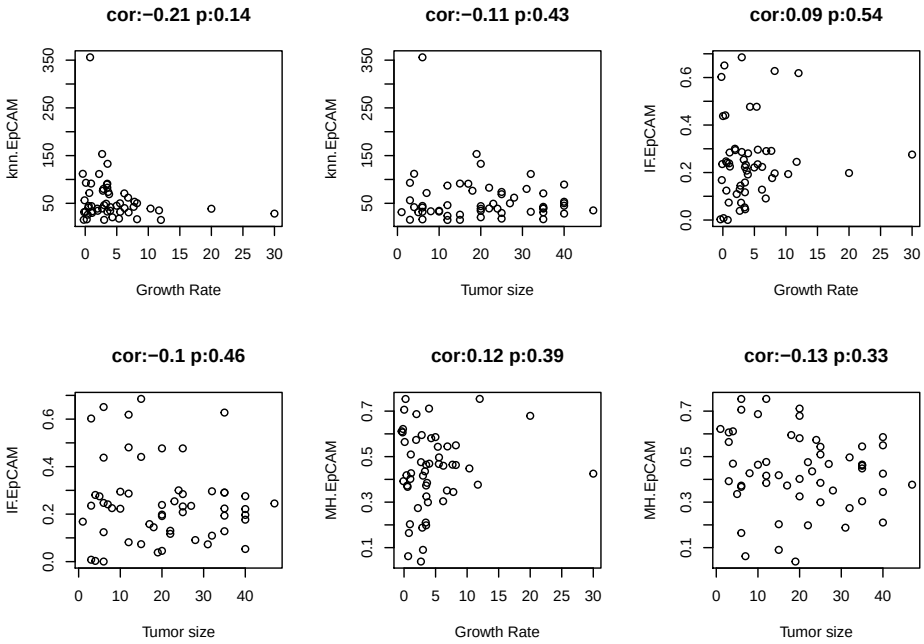
Firstly, we can compare the different metrics to determine how similar or different they are:



Do any of these metrics associate with growth or treatments?

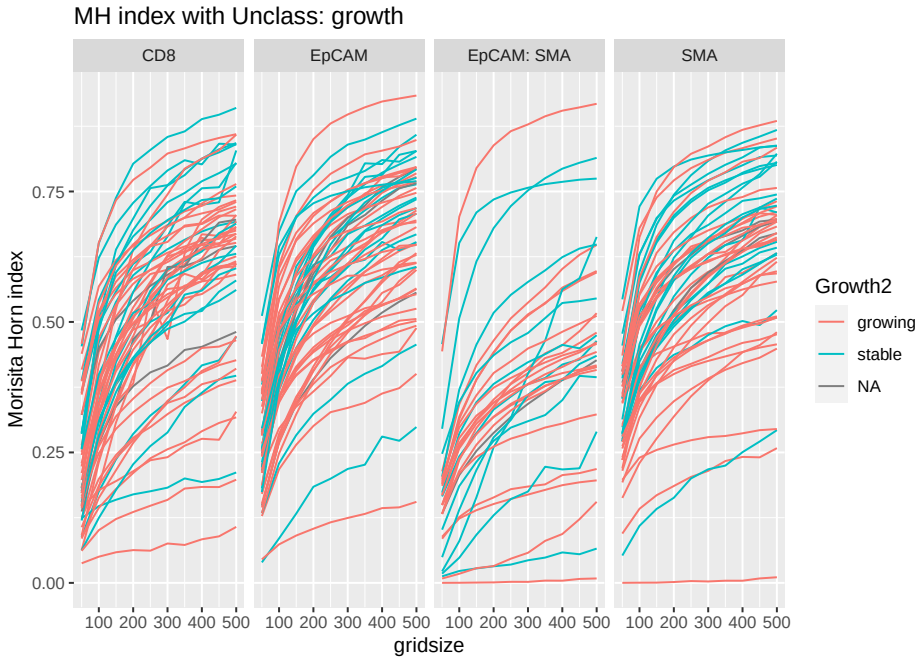


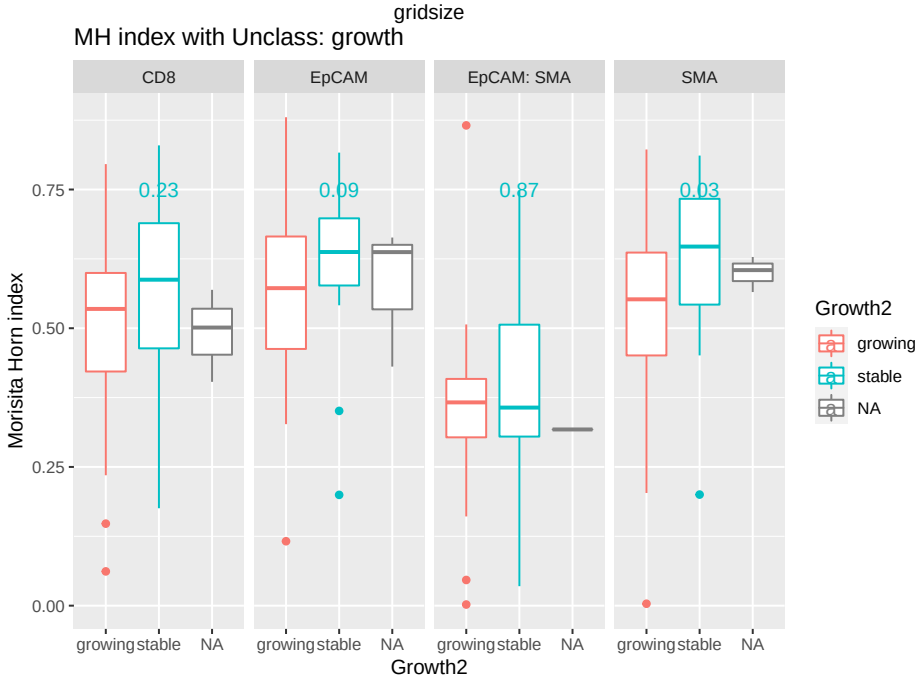
We can also compare this to raw tumor size or growth rate data as well:

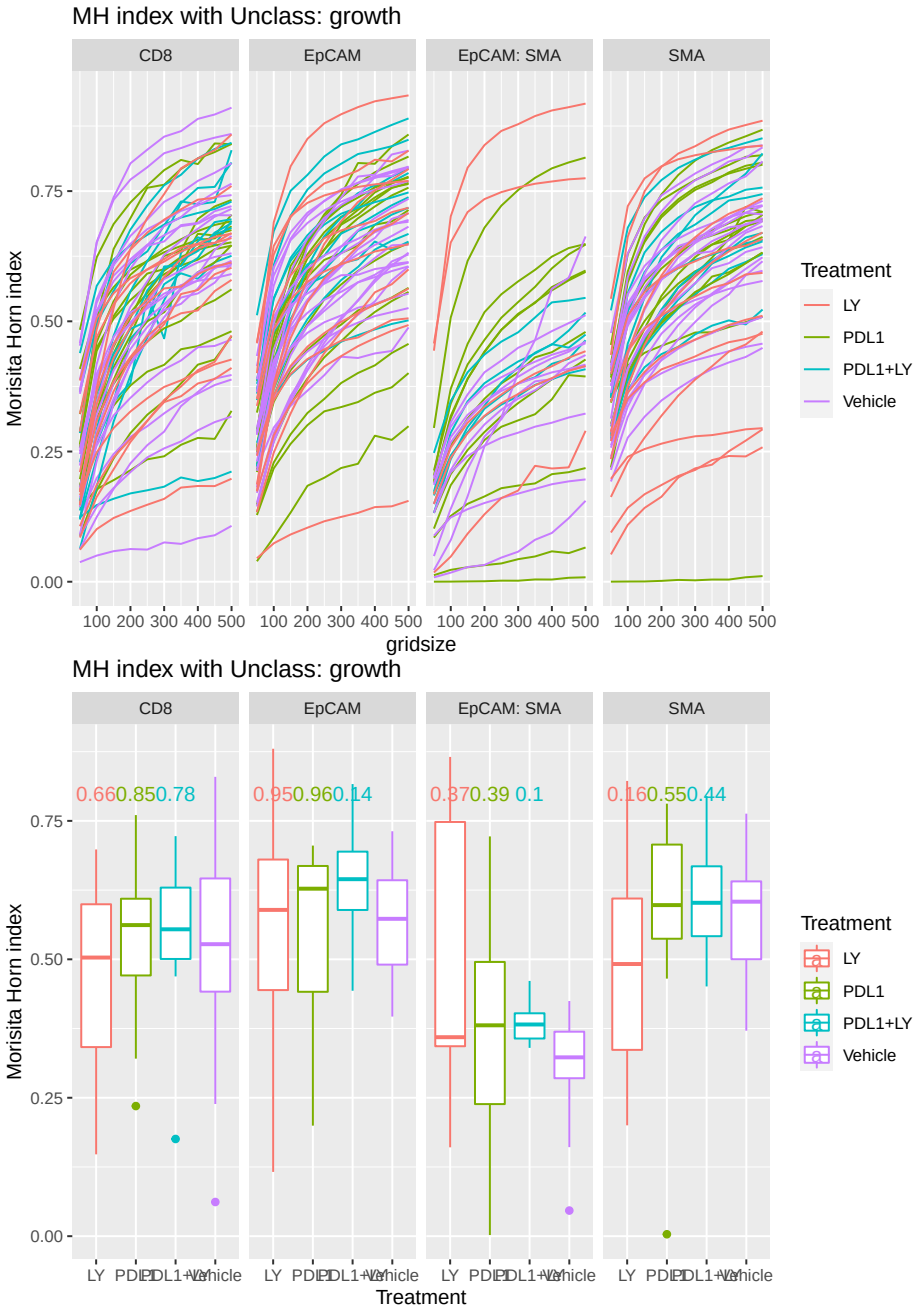


4.5 Other cell types

We noted that the proportion of Unclassified cells seemed to be different between the treatments. Assess here whether the MH index for this cell type is associated with growth or treatment here:


MH index with Unclass: growth
CD8
EpCAM
EpCAM: SMA
SMA
Morisita Horn index
0.00
0.25
0.50
0.75
100 200 300 400 500
gridsize
Growth2
growing
stable
NA


MH index with Unclass: growth
CD8
EpCAM
EpCAM: SMA
SMA
Morisita Horn index
0.00
0.25
0.50
0.75
0.23
0.09
0.87
0.03
growing stable NA
Growth2
Growth2
growing
stable
NA


MH index with Unclass: growth
CD8
EpCAM
EpCAM: SMA
SMA
Morisita Horn index
0.00
0.25
0.50
0.75
100 200 300 400 500
gridsize
Treatment
LY
PDL1
PDL1+LY
Vehicle
MH index with Unclass: growth
0.66 0.85 0.78
0.95 0.96 0.14
0.37 0.39 0.1
0.16 0.55 0.44
LY
PDL1
PDL1+LY
Vehicle
Treatment

Chapter 5

Expression data

This file looks at loading and pre-processing data for:

- differential gene expression analysis
- PAM50 subtyping
- uploading into CIBERSORT/TIMER

5.1 Running alignment

Samples were mapped in star using the following parameters. Note that the first two batches of samples run had shorter read lengths (~75 bp) whereas batch 3 had lengths of ~150bp

```
## Not run here
STAR \
     --readFilesCommand zcat \
     --genomeDir /n/scratch2/at268/rn6_v2 \
     --sjdbGTFfile /n/scratch2/at268/rn6_v2/rn6.refGene.gtf \
     --runThreadN 10 \
     --runMode alignReads \
     --genomeLoad NoSharedMemory\
     --outSAMattributes NH HI AS nM NM\
     --outSAMstrandField intronMotif\
     --outFilterMultimapNmax 20\
     --alignSJoverhangMin 8\
     --readFilesIn $1 $2 \
     --alignSJDBoverhangMin 1\
     --outFilterMismatchNmax 999\
     --outFilterMismatchNoverLmax 0.1\
     --alignIntronMin 20\
     --alignIntronMax 1000000\
```

```
    --alignMatesGapMax 1000000\
    --outFilterType BySJout\
    --outFilterScoreMinOverLread 0.33 \
    --outFilterMatchNminOverLread 0.33 \
    --limitSjdbInsertNsj 1200000 \
    --outFilterIntronMotifs None \
    --alignSoftClipAtReferenceEnds Yes\
    --outSAMattrRGline ID:$4 SM:$4 \
    --chimSegmentMin 15 \
    --chimJunctionOverhangMin 15\
    --limitBAMsortRAM 0\
    --outSAMtype BAM SortedByCoordinate\
    --outSAMunmapped Within \
    --quantMode GeneCounts transcriptomeSAM \
    --quantTranscriptomeBan IndelSoftclipSingleend \
    --outFileNamePrefix $3 \
    --twopassMode Basic
```

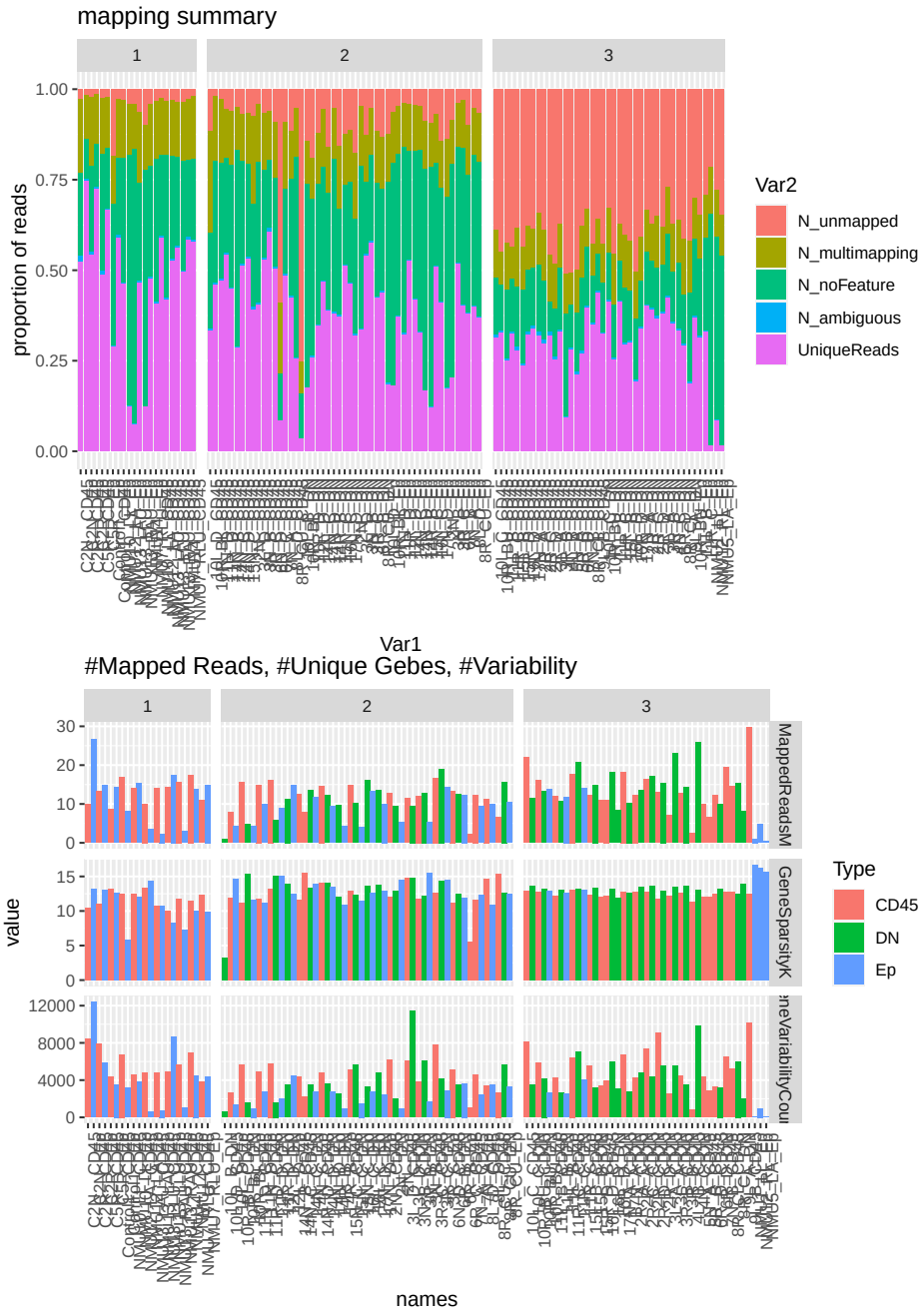
In addition, RSEM was run to obtain TPM, rsem and FPKM counts. Note: rn6.refGene.gtf.gz was used to generate the RSEM library!(This is the may version)

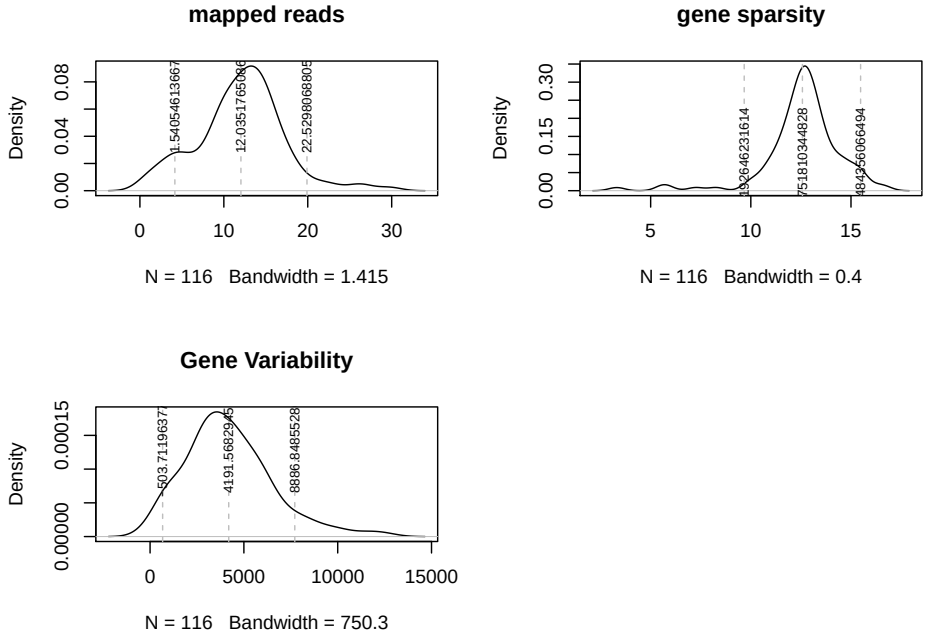
5.2 RNA Initial QC

Default output from R showing the number of unique reads compared to multimapped, unmapped etc. This is shown for each batch. Note that batch 3 has differences (high percentage of unmapped) compared to the other batches, possibly due to DNA contamination.

Below we check for three measures:

- mapped million reads (ideally, 10M+ reads)
- Gene Sparsity: This is a measurement of the number of genes which have non-zero values. Ideally, would be greater than 10K, but values which are too high may also suggest contamination from DNA (unexpressed genes are also counted)
- Varability: standard deviation of the transcriptomic counts. If this value is too low, would suggest that high DNA contamination, non-representative transcriptome.


mapping summary
1
2
3
1.00
0.75
0.50
0.25
0.00
proportion of reads
Var2
N_unmapped
N_multimapping
N_noFeature
N_ambiguous
UniqueReads
Var1
#Mapped Reads, #Unique Gebes, #Variability
MappedReadsM
GeneSparsityK
neVariabilityCou
value
Type
CD45
DN
Ep
names



Samples to remove from analysis:

The thresholds indicated below are based on the above density plots, and removes cases which are <1.5 SD of the mean

- low total number of mapped reads (under 1.5M)
- sparsity: less than 8K genes
- variability : threshold under 500

The omitted samples are:

	MappedReadsM	GeneSparsityK	Batch	GeneVariabilityCounts	Type	r
6R_C_CD45	2.287110	5.585	2	1063.38925	CD45	6
10L_B_DN	0.864838	3.308	2	559.30659	DN	1
NMU1_LL_Ep	1.032440	16.708	3	69.82645	Ep	N
NMU5_LA_Ep	0.500429	15.707	3	35.56084	Ep	N
Control1___Ep	8.224593	5.800	1	3164.00515	Ep	C
NMU13_RAU_Ep	3.049733	7.242	1	1069.99912	Ep	N

We are left with 110 samples.

There are 47, 32, 31 samples in the CD45, Ep, DN fractions.

There are 20, 49, 41 samples from batches 1, 2 and 3 respectively.

5.3 Normalisation

Run through DESEq and normalise the library. Using all samples, we run the model:

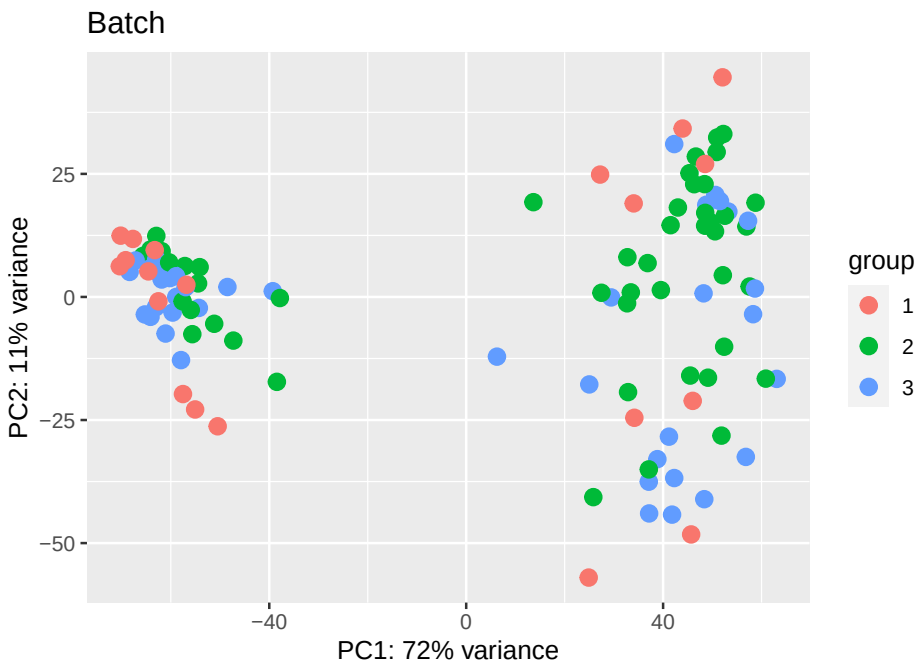
expression ~ Celltype + factor (Batch)

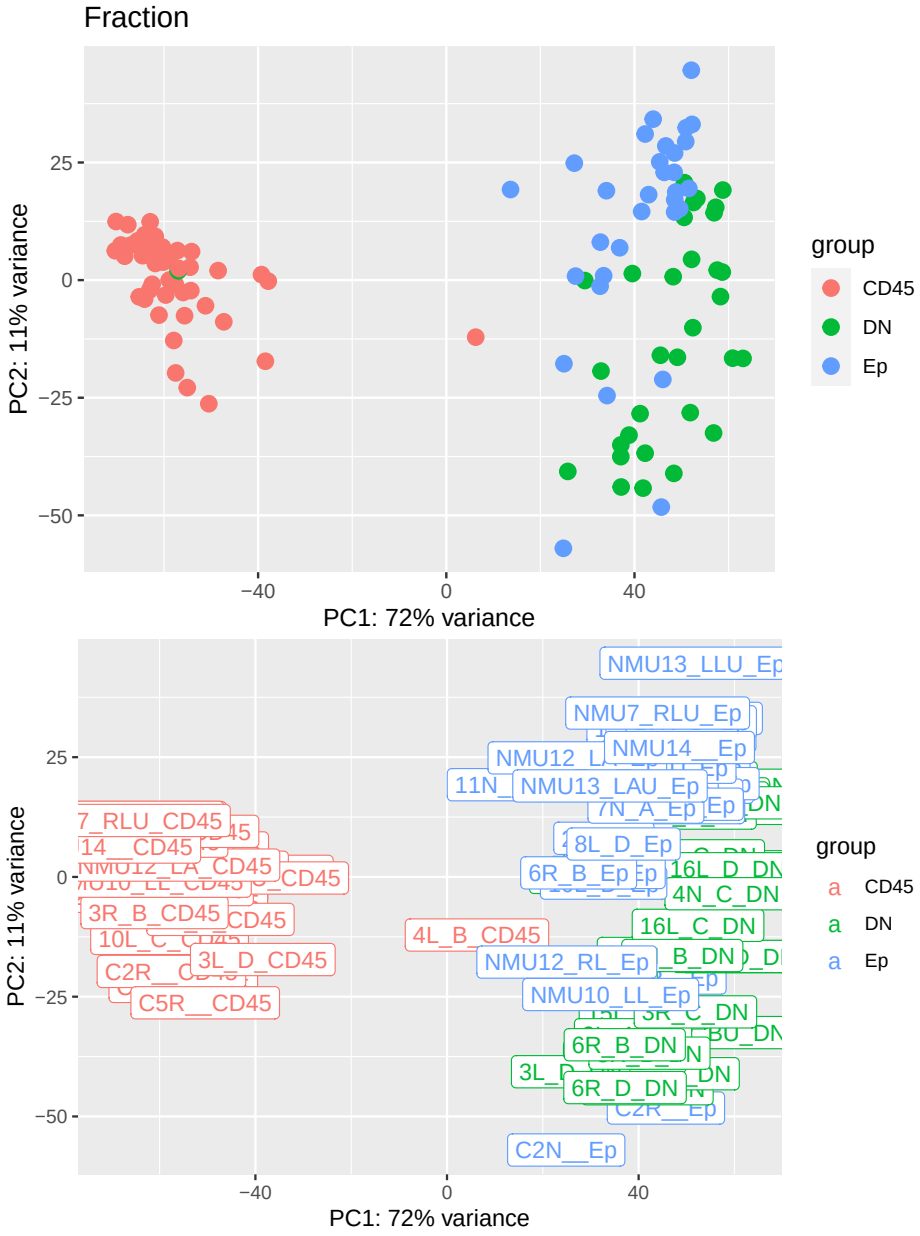
and keep the genes which have a total count of at least half the number of samples. ie. $ sum(gene_i) > N_{samples}/2 $

5.3.1 preliminary visualisation (to remove outliers)

Below are PCA plots based on:

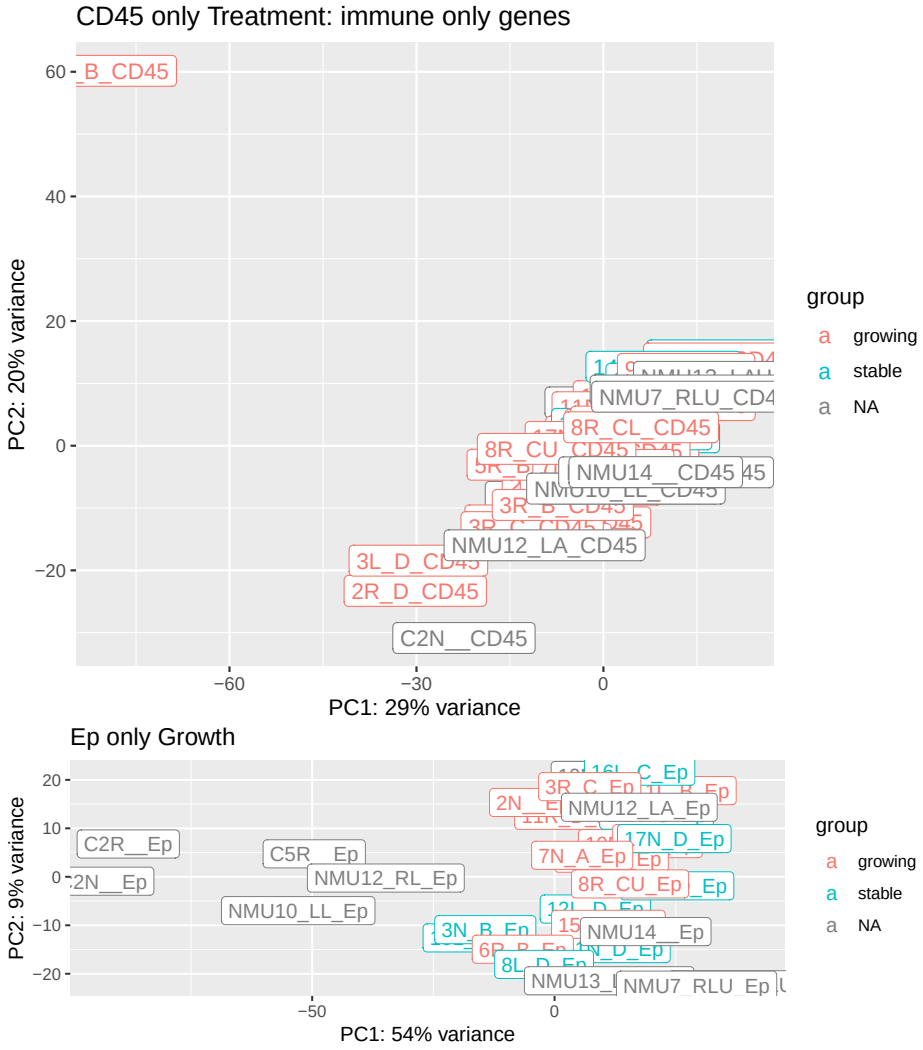
- Batch
- CellType

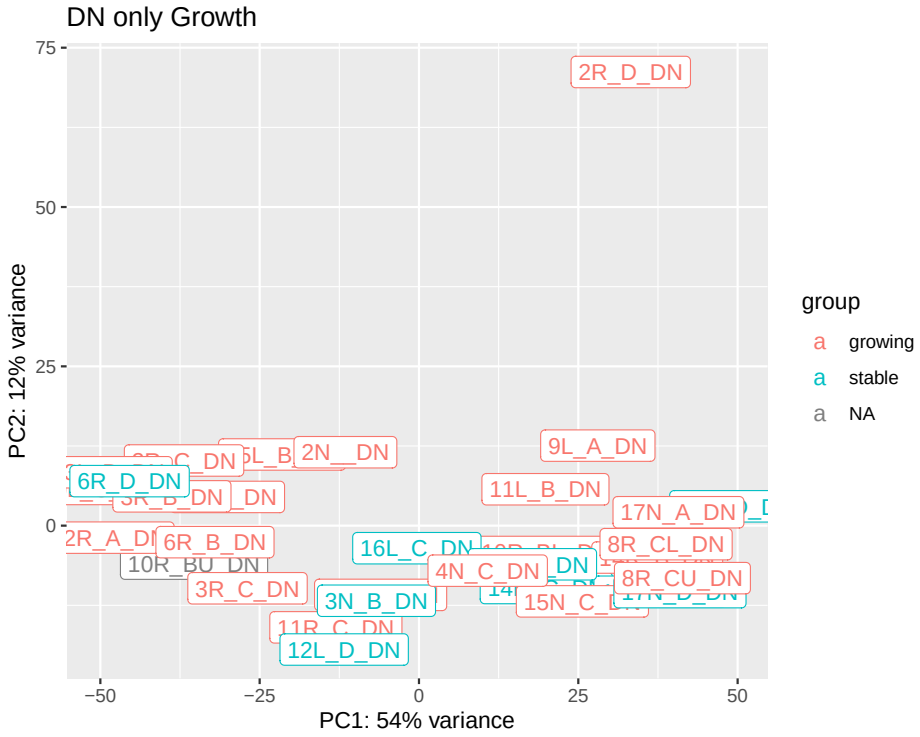




Batches in general separate out well, however, some samples appear to be outliers in comparison to the main group. We look in closer detail the CD45, DN and EpCAM populations.

In the CD45 population, narrow down to only immune related genes to see if there is a difference.


CD45 only Treatment: immune only genes
PC2: 20% variance
PC1: 29% variance
60
40
20
0
−20
−60
−30
0
_B_CD45
NMU7_RLU_CD4
8R_CL_CD45
8R_CU_CD45
NMU14__CD45
NMU12_LA_CD45
3L_D_CD45
2R_D_CD45
C2N__CD45
group
a growing
a stable
a NA
Ep only Growth
PC2: 9% variance
PC1: 54% variance
20
10
0
−10
−20
−50
0
C2R__Ep
2N__Ep
C5R__Ep
NMU12_RL_Ep
NMU10_LL_Ep
NMU12_LA_Ep
17N_D_Ep
7N_A_Ep
8R_CU_Ep
3N_B_Ep
NMU14__Ep
NMU7_RLU_Ep
group
a growing
a stable
a NA



Based on the above plots, we remove the following outliers and re-run the normalisation:

- 2R_D_DN
- 4L_B_CD45

5.4 Processing files for external software

We also process these files for external software:

Software	RNA-data
PAM50	rsem data after quantile normalised, genes converted to human
TIMER	TPM values, retain rat ids
xcell	rsem data, genes converted ti human
cibersort	tpm?? convert to rat genes and use lm22 rat

```
## named integer(0)
```

```
## named integer(0)
```

Save the mouse names for TIMER cistrome: check that this is actually required for TIMER

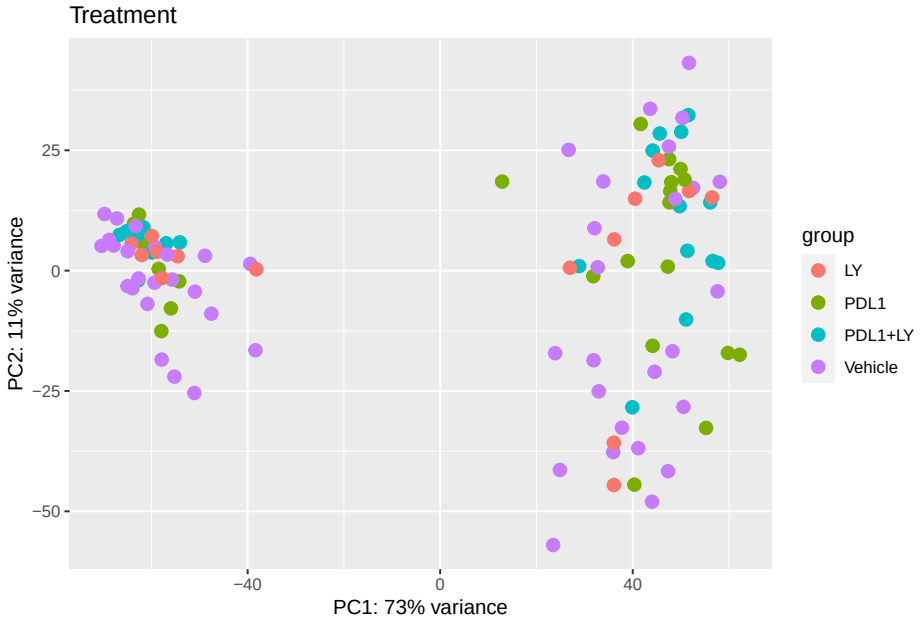
Chapter 6

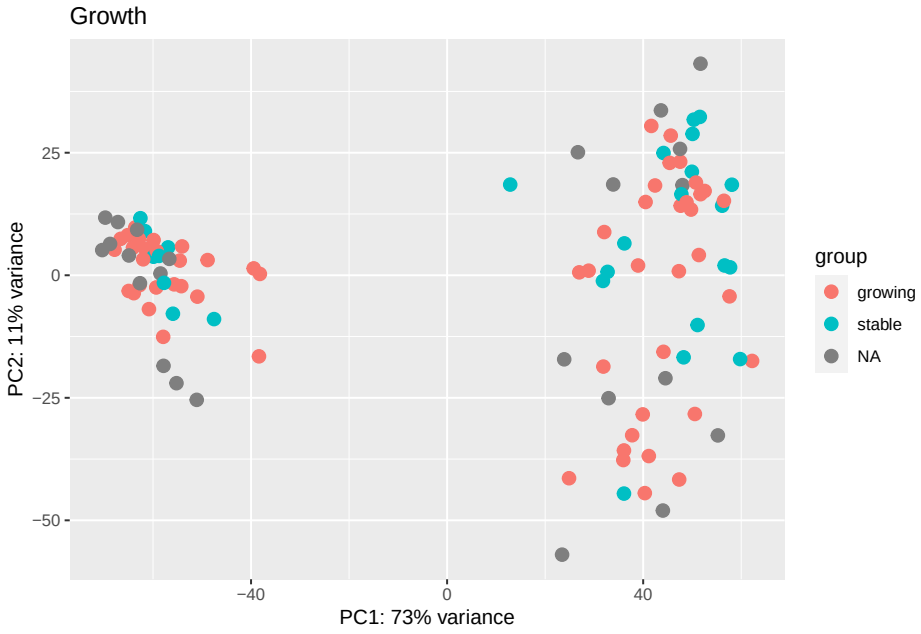
RNA data: preliminary plots

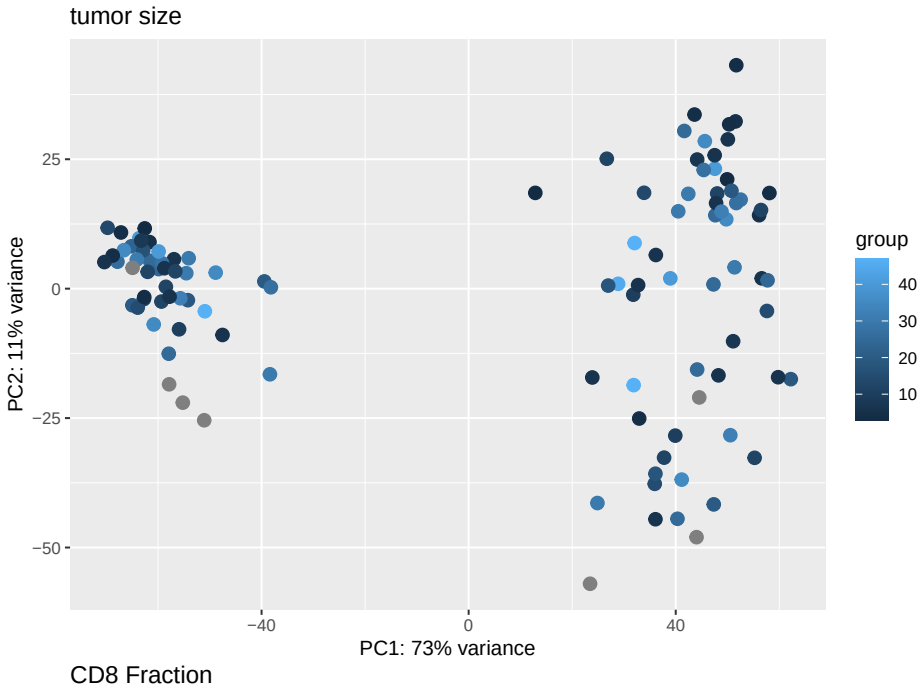
Prior to doing any comparative analysis, we will look at the following plots to get an overview of the data.

6.1 PCA plots

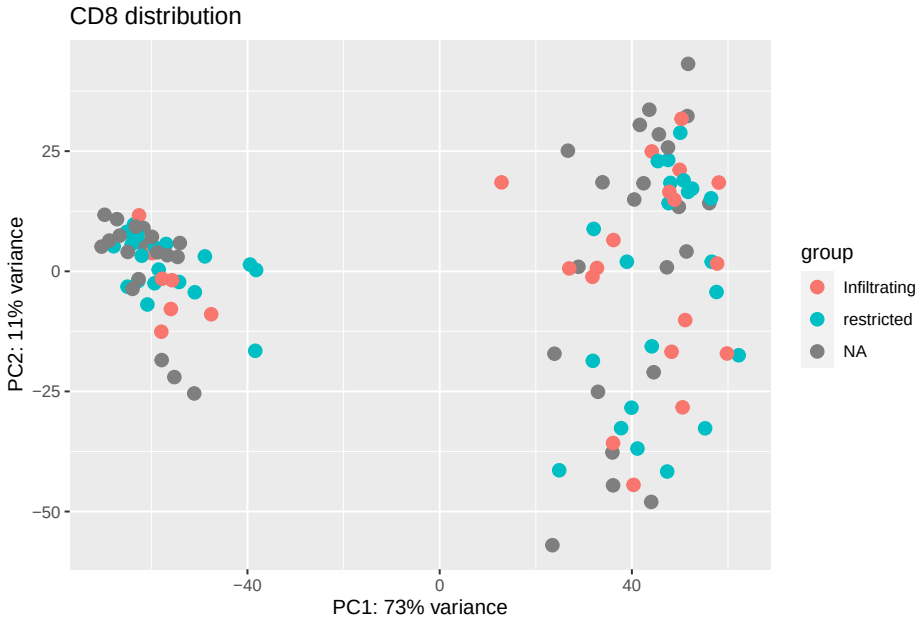
We can check the new PCA plots, and overlay parameters of interest including treatment, growth, tumor size, CD8 fraction, spatial distribution.

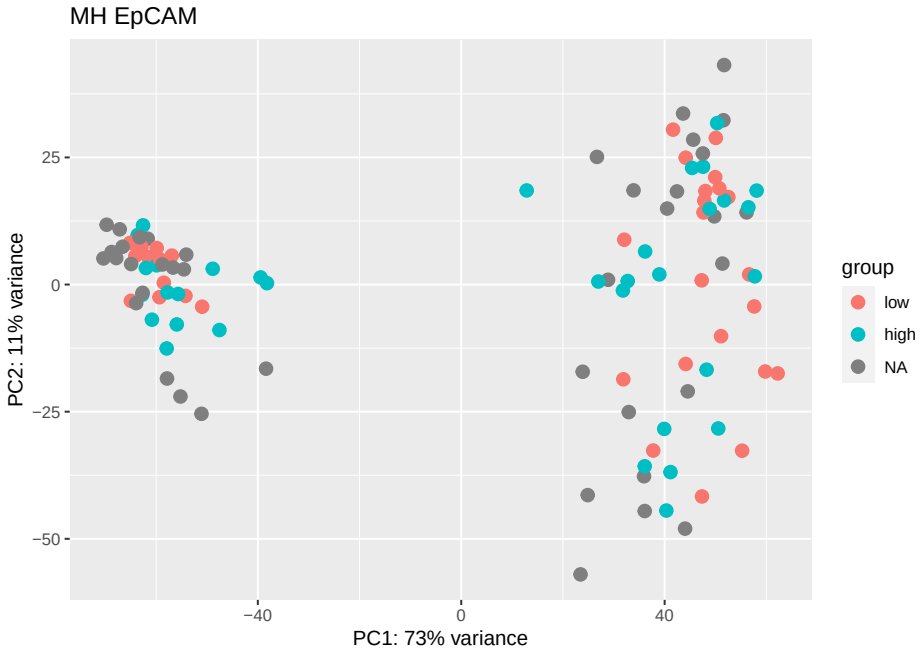

Treatment
25
0
−25
−50
PC2: 11% variance
−40
0
40
PC1: 73% variance
group
LY
PDL1
PDL1+LY
Vehicle

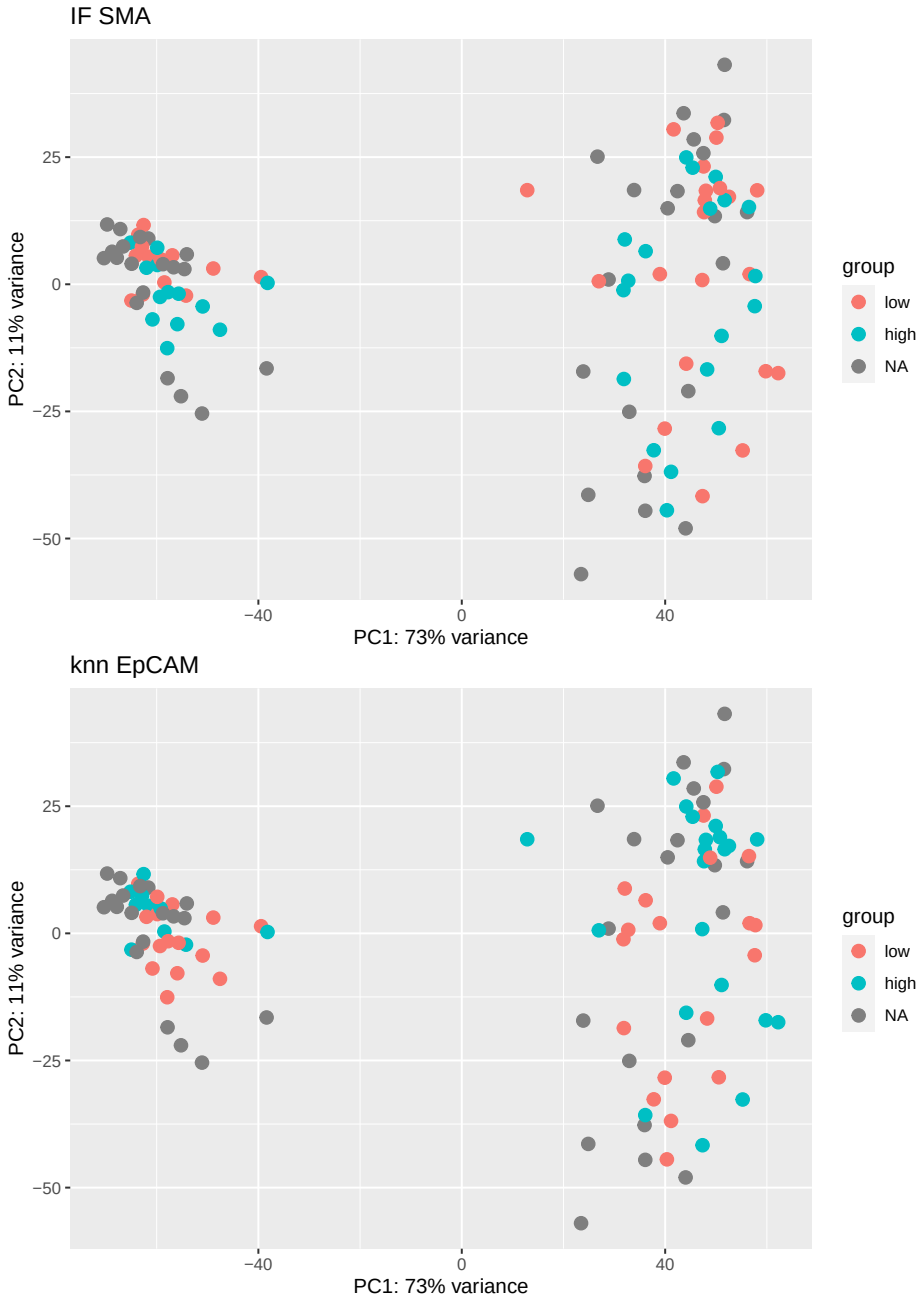

Growth
25
0
−25
−50
PC2: 11% variance
−40
0
40
PC1: 73% variance
group
growing
stable
NA


tumor size
PC2: 11% variance
PC1: 73% variance
group
40
30
20
10
25
0
−25
−50
−40
0
40

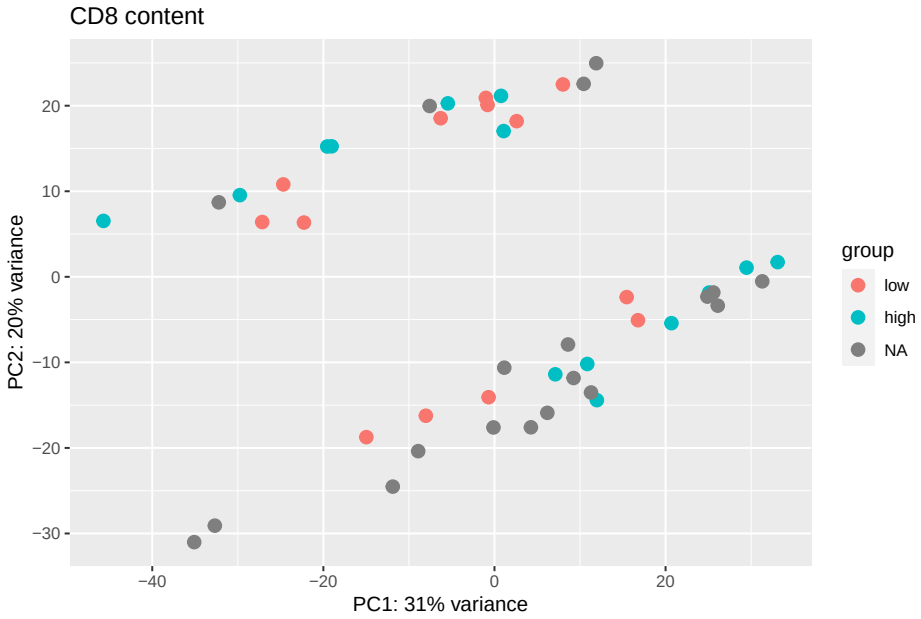

CD8 Fraction
PC2: 11% variance
PC1: 73% variance
group
low
high
NA
25
0
−25
−50
−40
0
40

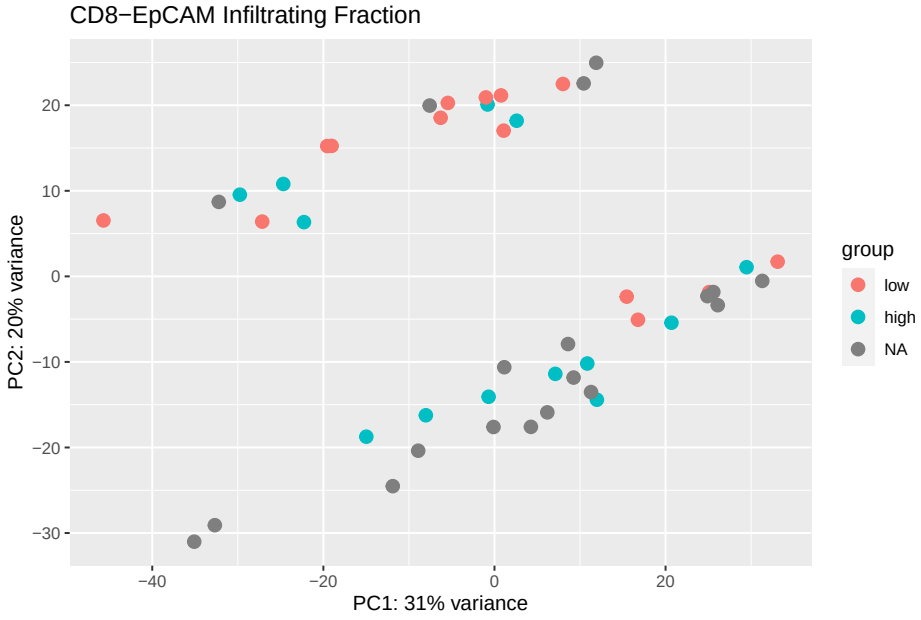

CD8 distribution
PC2: 11% variance
PC1: 73% variance
25
0
−25
−50
−40
0
40
group
Infiltrating
restricted
NA

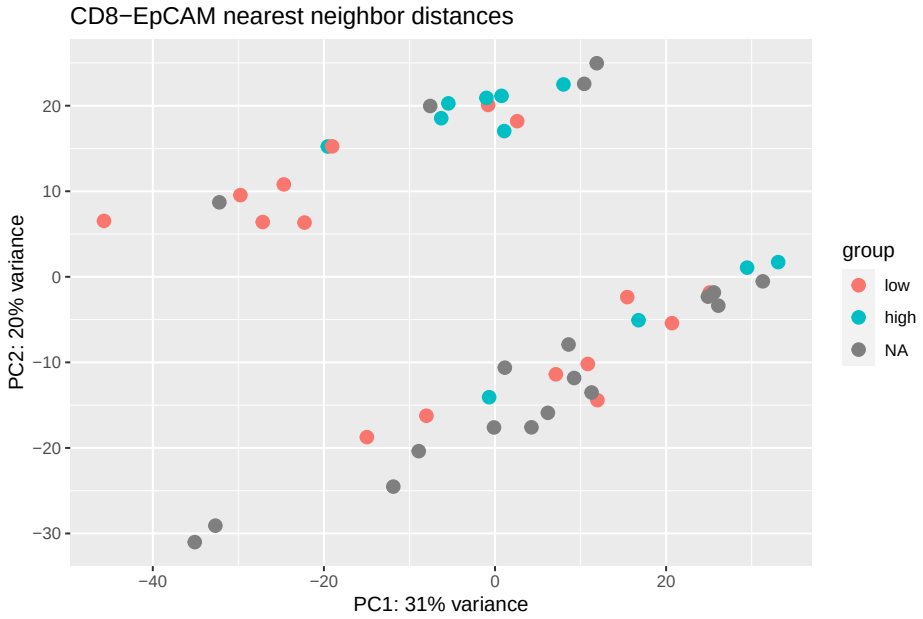

MH EpCAM
PC2: 11% variance
PC1: 73% variance
25
0
−25
−50
−40
0
40
group
low
high
NA

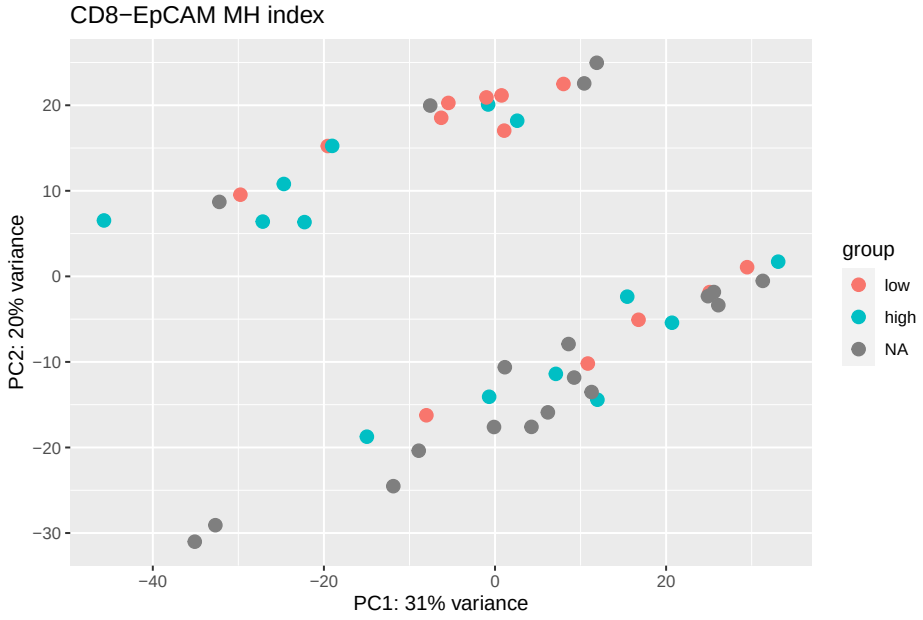


In addition, we can look at the CD45 population and the distributions based on CD8 content and spatial infiltration


CD8 content
PC2: 20% variance
PC1: 31% variance
group
low
high
NA


CD8−EpCAM Infiltrating Fraction
PC2: 20% variance
PC1: 31% variance
group
low
high
NA





6.2 Expression patterns by cell type

Below, we check whether the different fractions are expressing expected markers

The cell types are:

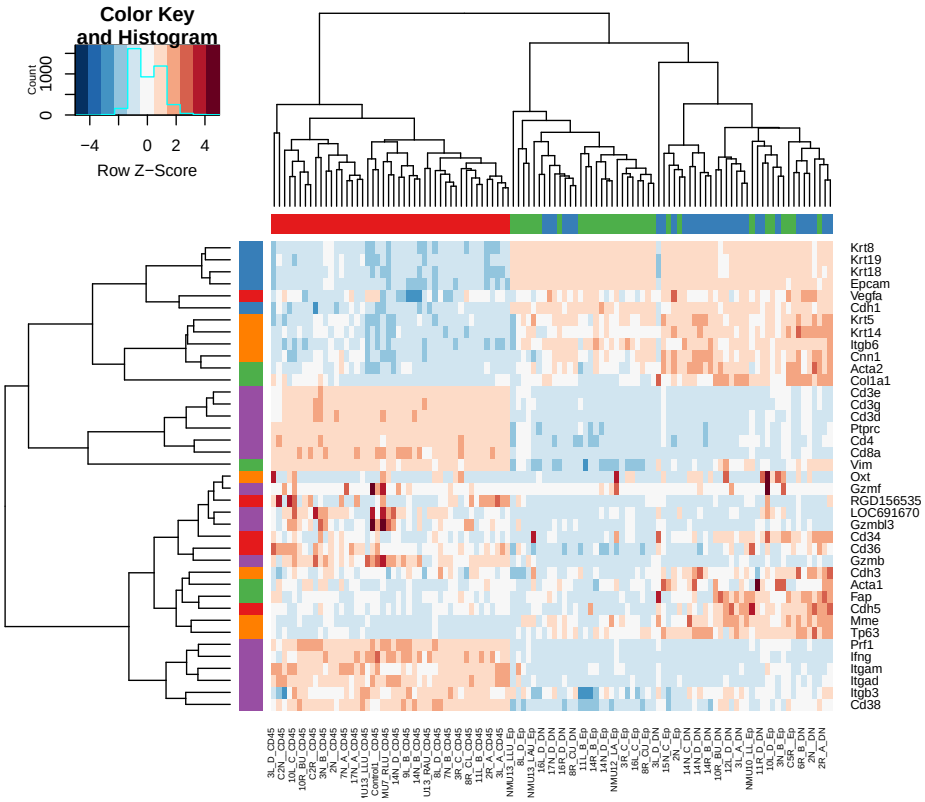
- Red: Cd45
- Epcam: green
- DN:blue

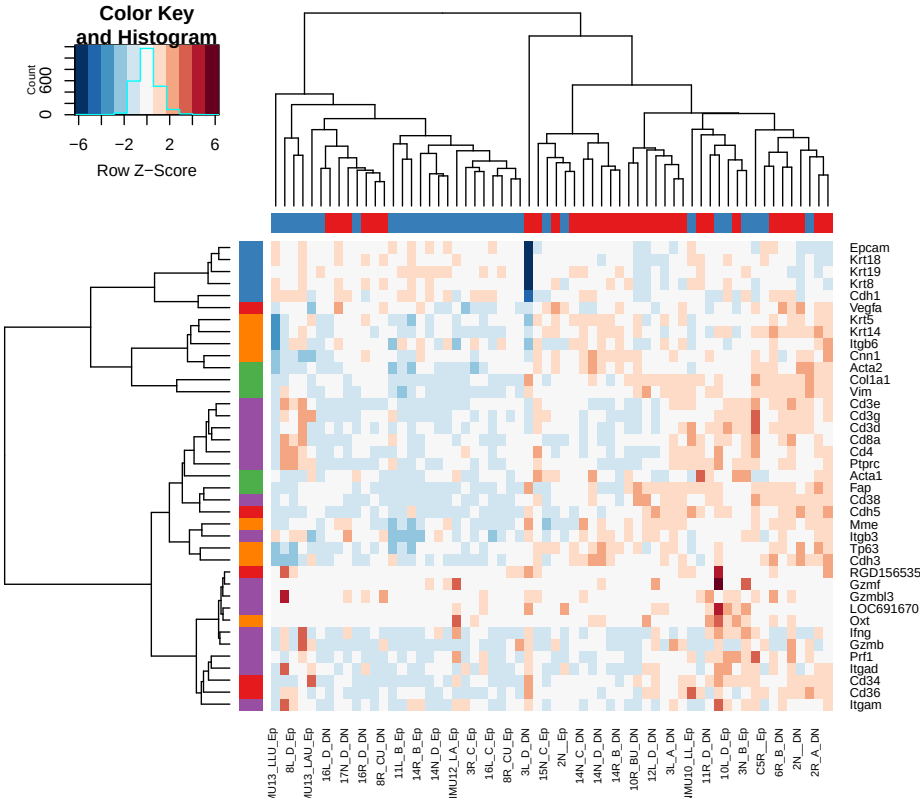
Reference Genes:

- purple: immune
- blue: epithelial
- green: stroma
- orange: myoepithelial
- red: endothelial

Below, we see that the CD45 cells separate from the Ep/DN populations have high expression of immune related genes including CD3, CD4, IFNG.

However, the DN/Ep fractions are more intermixed. The DN fraction has expression of keratins, as well as fibroblast markers (Acta1), and myeoepithelial markers (Tp63)




Color Key
and Histogram
Count
600
0
−6
−2
2
6
Row Z−Score
Epcam
Krt18
Krt19
Krt8
Cdh1
Vegfa
Krt5
Krt14
Itgb6
Cnn1
Acta2
Col1a1
Vim
Cd3e
Cd3g
Cd3d
Cd8a
Cd4
Ptprc
Acta1
Fap
Cd38
Cdh5
Mme
Itgb3
Tp63
Cdh3
RGD156535
Gzmf
Gzmbl3
LOC691670
Oxt
Ifng
Gzmb
Prf1
Itgad
Cd34
Cd36
Itgam
8L_D_Ep
16L_D_DN
17N_D_DN
16R_D_DN
8R_CU_DN
11L_B_Ep
14R_B_Ep
14N_D_Ep
3R_C_Ep
16L_C_Ep
8R_CU_Ep
3L_D_DN
15N_C_Ep
2N__Ep
14N_C_DN
14N_D_DN
14R_B_DN
10R_BU_DN
12L_D_DN
3L_A_DN
11R_D_DN
10L_D_Ep
3N_B_Ep
C5R__Ep
6R_B_DN
2N__DN
2R_A_DN

Chapter 7

DESeq analysis

This document sets up DESeq runs to compare:

- DN vs Ep samples
- growing vs stable samples (all 3 fractions)
- treatments (all 3 fractions)
- spatial patterns (all 3 fractions)

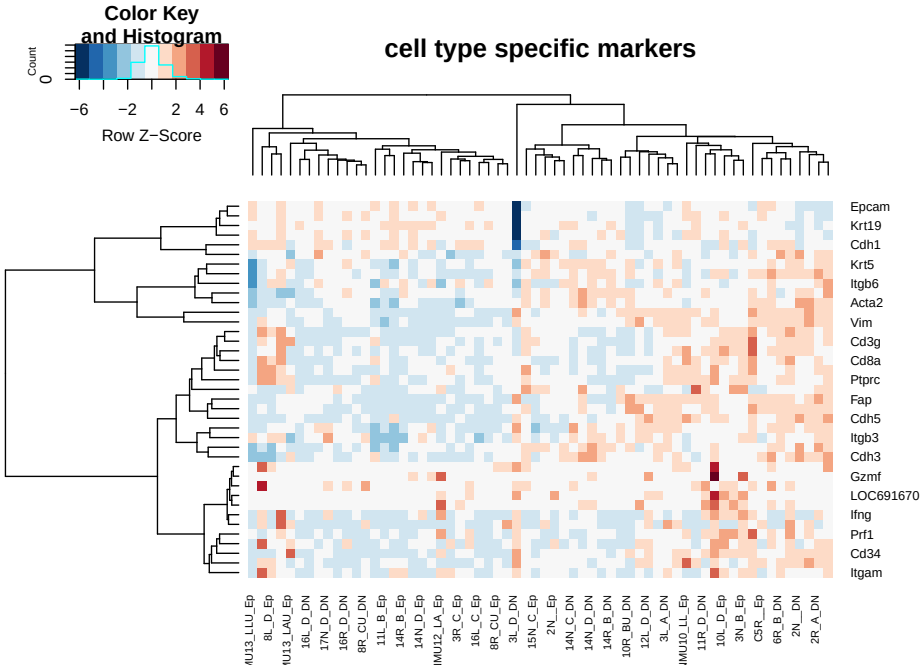
The outputs of these analyses will be used for Gene Set Enrichment Analysis, using MSigDB databases (c2, c5, hallmark), alongside pathways from Metacore (Process Networks, Pathway Maps)

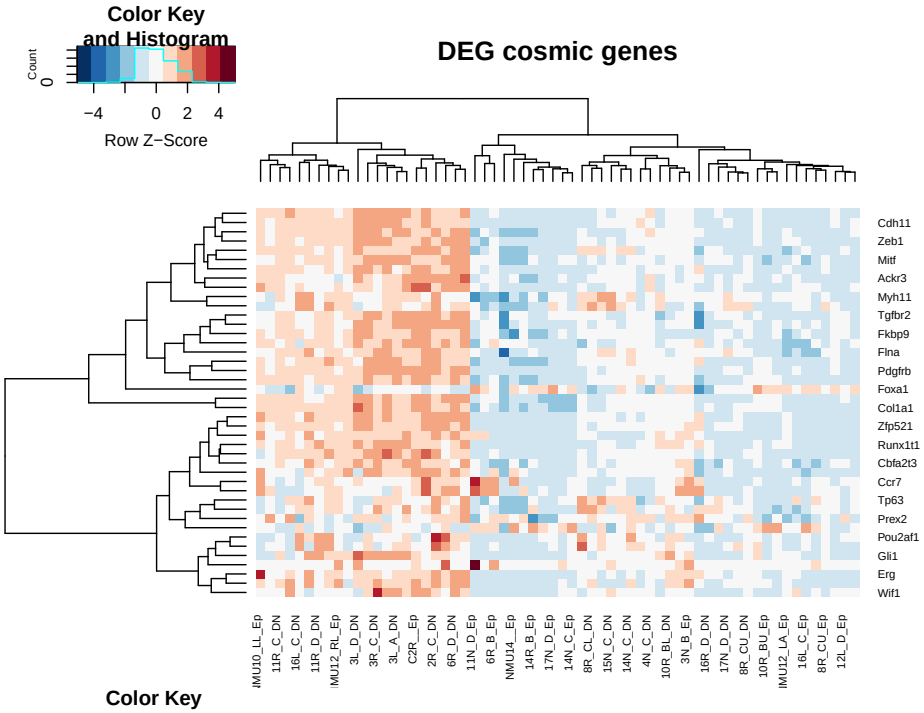
7.1 DN vs Ep

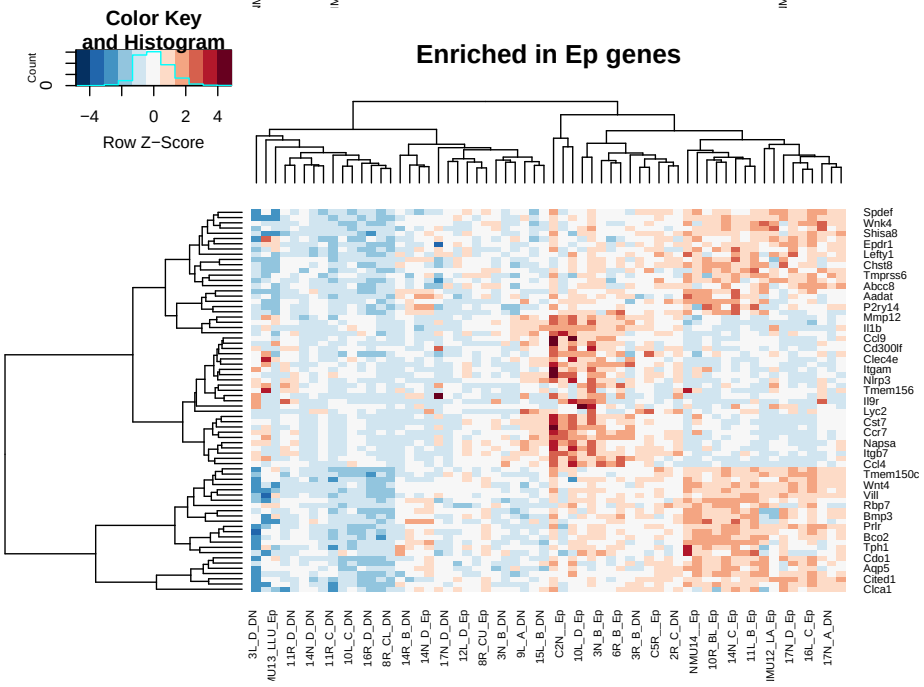
What is the difference between DN and Ep fractions? Select for genes with p value cut off of 0.05 and log2 fold change of 1.5. Also, when visualising top genes, select for base expression greater than 100

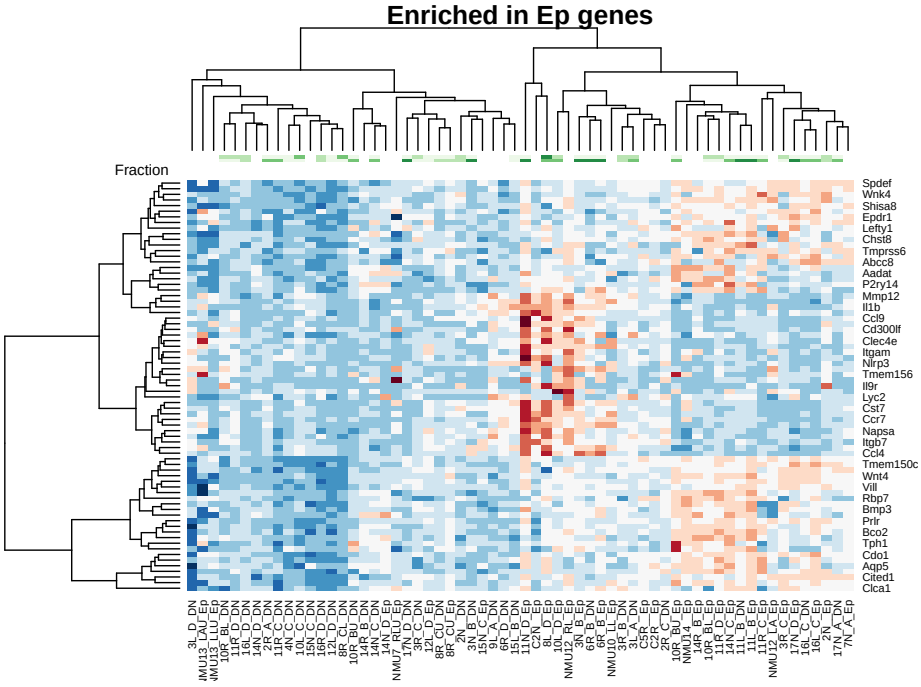
```
##
## out of 15301 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 2172, 14%
## LFC < 0 (down)     : 1188, 7.8%
## outliers [1]       : 463, 3%
## low counts [2]     : 779, 5.1%
## (mean count < 2)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
Fap	147.505781	3.791943	0.6564706	5.776257	0.0000000	0.0000009
Vim	13706.172521	1.832658	0.3361022	5.452680	0.0000000	0.0000035
RGD1565355	4.841686	-2.427252	0.8547683	-2.839661	0.0045162	0.0307470
Mme	214.622521	2.418161	0.5763311	4.195784	0.0000272	0.0005517
Cnn1	618.403495	1.572872	0.4698537	3.347578	0.0008152	0.0081399
Tp63	145.596811	2.047668	0.4768123	4.294494	0.0000175	0.0003907
Cdh5	114.363936	1.899697	0.6674752	2.846093	0.0044259	0.0302941
Itgam	289.807818	-2.251148	0.5564127	-4.045824	0.0000521	0.0009532
Col1a1	35939.383628	4.962384	0.7227443	6.866031	0.0000000	0.0000000
Krt14	949.200864	1.834436	0.4774425	3.842214	0.0001219	0.0018837
Acta2	489.516373	1.678424	0.4013367	4.182086	0.0000289	0.0005810
Cd38	370.418350	1.897076	0.4719555	4.019608	0.0000583	0.0010370
Cdh3	100.102955	1.705776	0.4293903	3.972553	0.0000711	0.0012132




Color Key and Histogram
Count
0
−4
0
2
4
Row Z−Score
DEG cosmic genes
Cdh11
Zeb1
Mitf
Ackr3
Myh11
Tgfbr2
Fkbp9
Flna
Pdgfrb
Foxa1
Col1a1
Zfp521
Runx1t1
Cbfa2t3
Ccr7
Tp63
Prex2
Pou2af1
Gli1
Erg
Wif1
IMU10_LL_Ep
11R_C_DN
16L_C_DN
11R_D_DN
IMU12_RL_Ep
3L_D_DN
3R_C_DN
3L_A_DN
C2R_Ep
2R_C_DN
6R_D_DN
11N_D_Ep
6R_B_Ep
NMU14_Ep
14R_B_Ep
17N_D_Ep
14N_C_Ep
8R_CL_DN
15N_C_DN
14N_C_DN
4N_C_DN
10R_BL_DN
3N_B_Ep
16R_D_DN
17N_D_DN
8R_CU_DN
10R_BU_Ep
IMU12_LA_Ep
16L_C_Ep
8R_CU_Ep
12L_D_Ep


Color Key and Histogram
Count
0
−4
0
2
4
Row Z−Score
Enriched in Ep genes
Spdef
Wnk4
Shisa8
Epdr1
Lefty1
Chst8
Tmprss6
Abcc8
Aadat
P2ry14
Mmp12
Il1b
Col9
Cd300lf
Clec4e
Itgam
Nlrp3
Tmem156
Il9r
Lyc2
Cst7
Ccr7
Napsa
Itgb7
Ccl4
Tmem150c
Wnt4
Vill
Rbp7
Bmp3
Prlr
Bco2
Tph1
Cdo1
Aqp5
Cited1
Clca1
3L_D_DN
IMU13_LLU_Ep
11R_D_DN
14N_D_DN
11R_C_DN
10L_C_DN
16R_D_DN
8R_CL_DN
14R_B_DN
14N_D_Ep
17N_D_DN
12L_D_Ep
8R_CU_Ep
3N_B_DN
9L_A_DN
15L_B_DN
C2N__Ep
10L_D_Ep
3N_B_Ep
6R_B_Ep
3R_B_DN
C5R__Ep
2R_C_DN
NMU14__Ep
10R_BL_Ep
14N_C_Ep
11L_B_Ep
IMU12_LA_Ep
17N_D_Ep
16L_C_Ep
17N_A_DN



Note that there seems to be two main groups of Ep cells:

The first one is enriched for canonical tumour progression pathways: eg. expression of wnt, fox1a etc. The second one doesn't seem to have as high expression of these genes, but have high immune related signalling pathways

7.2 Set-up the comparisons

Separate out the tables into CD45, DN and Ep fractions. Compare

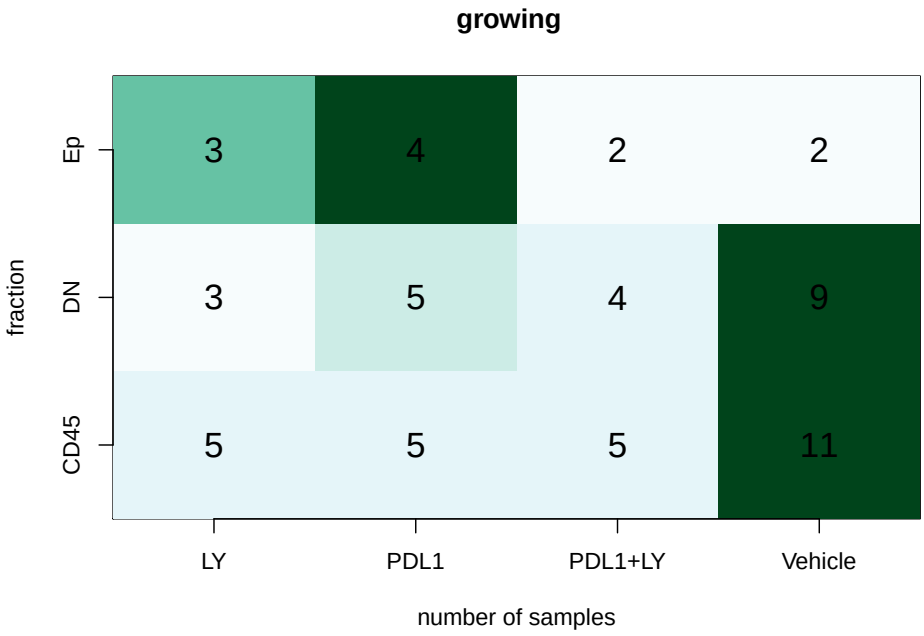
1. Differences based on treatment (and treatment vs control)
2. Specific treatments vs control
3. Growing vs stable (overall)
4. Growing vs stable (in each subgroup)

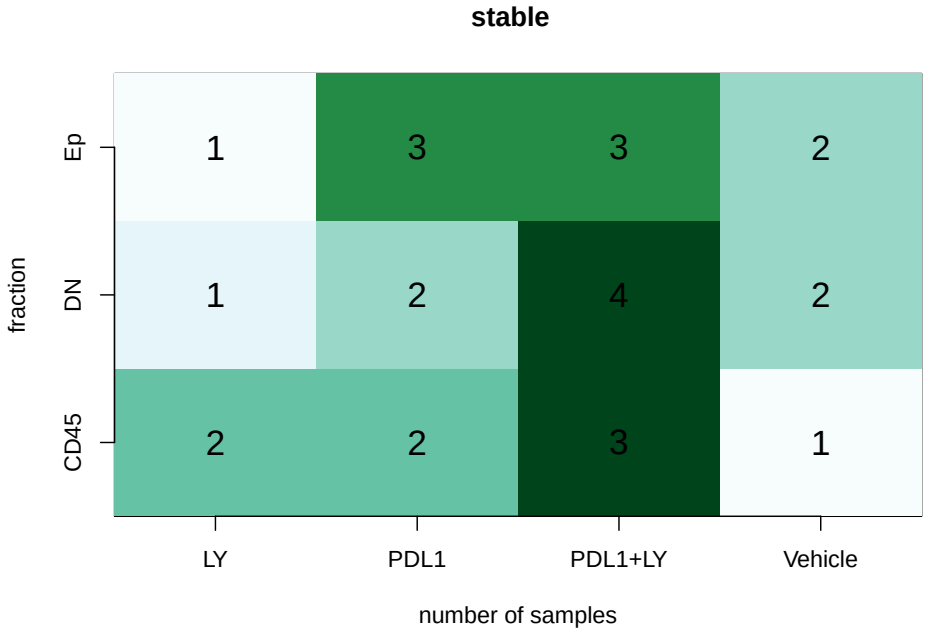
Check there are enough samples in each subgroup:

```
## [1] "number of samples for each treatment and growth rate"
```

```
## , ,  = growing
##
##
##        LY PDL1 PDL1+LY Vehicle
##   CD45  5    5       5      11
##   DN    3    5       4       9
##   Ep    3    4       2       2
```

```
##
## , ,  = stable
##
##
##        LY PDL1 PDL1+LY Vehicle
##   CD45  2    2       3       1
##   DN    1    2       4       2
##   Ep    1    3       3       2
```





Setup the following comparisons for each cell type:

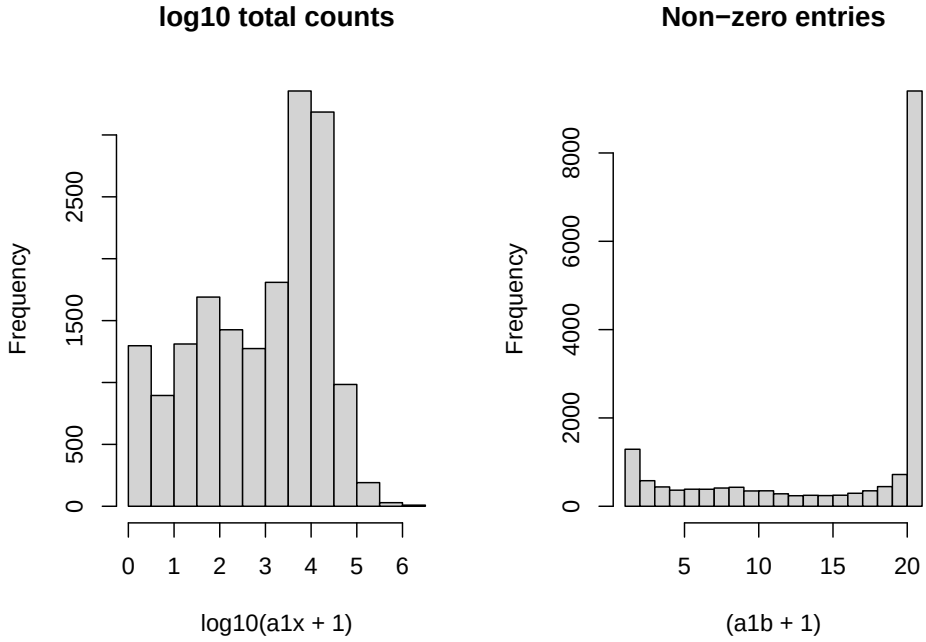
expression ~ growing + treatment + batch (exclude 1) (start with vehicle vs some treatment)

4 categories only: immunotherapy vs non-immuno

Separate out the tables into CD45, DN and Ep fractions. Compare:

1. Treatment (EpddsTreat)
2. Comp4: Growth + any immunotherapy treatment (Epdds)
3. Comp8: Growth + immunotherapy (EpddsTreatG)
4. Growth alone: (EpddsGrowth)
5. MH index: (EpddsStrMH)
6. Interacting Fraction (EpddsStrIF)
7. knn values (EpddsStrknn)
8. CD8 content (EpddsCD8)
9. Manual scoring spatial (EpddsSpatMan)
10. MH index + CD8 frac (Epddscd8MH)

We remove genes which have 0 counts in more than half the samples, and genes which have a row sum less than 10^(log10(mean(rowsums))-log10(sd(rowsums)))

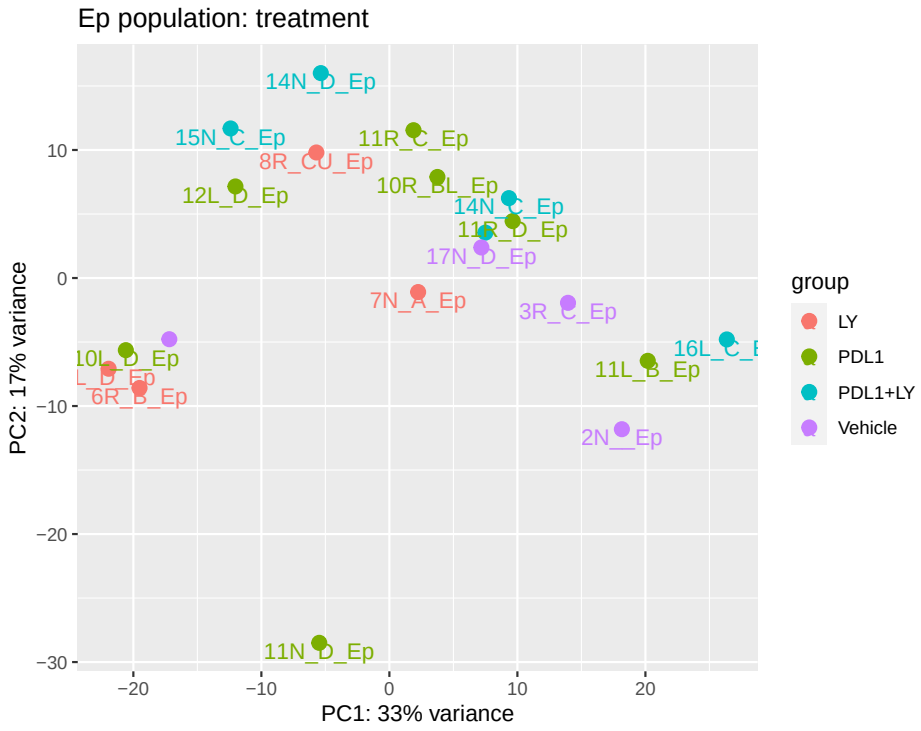


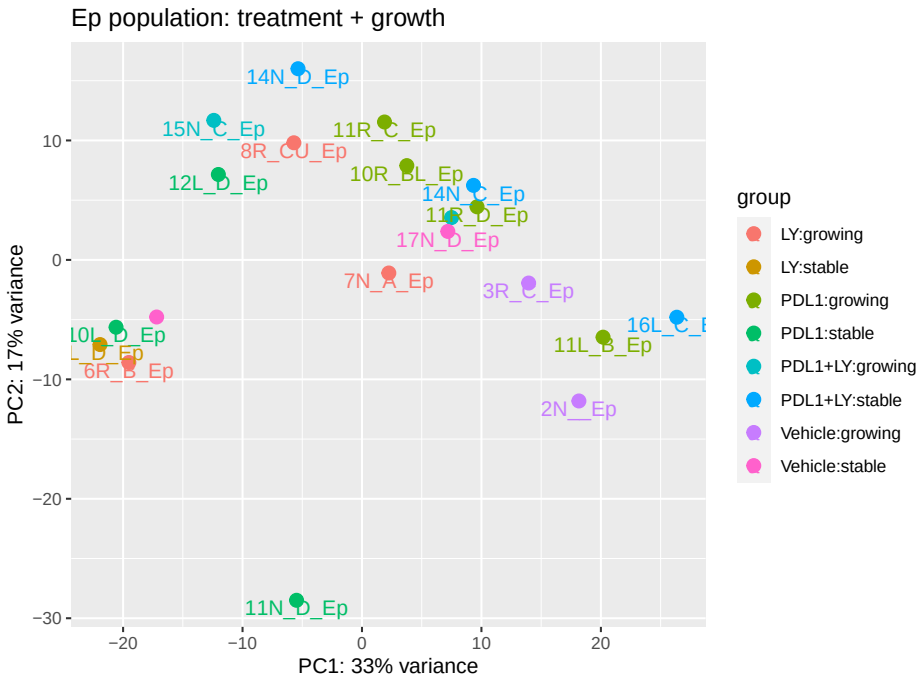
temp insertion: do comparison of differences in spatial with outcome:

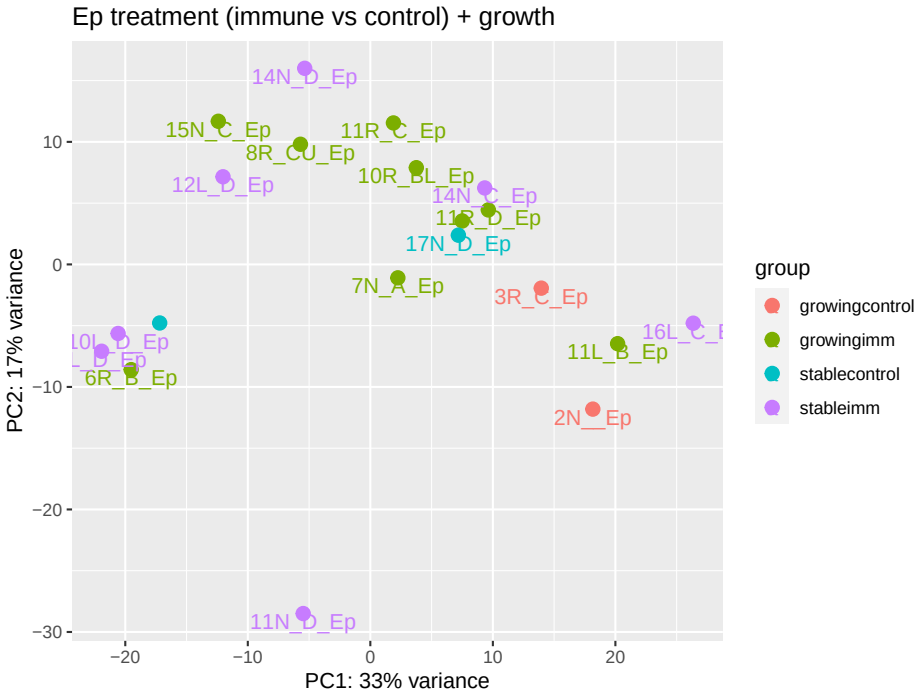
7.3 PCA plots

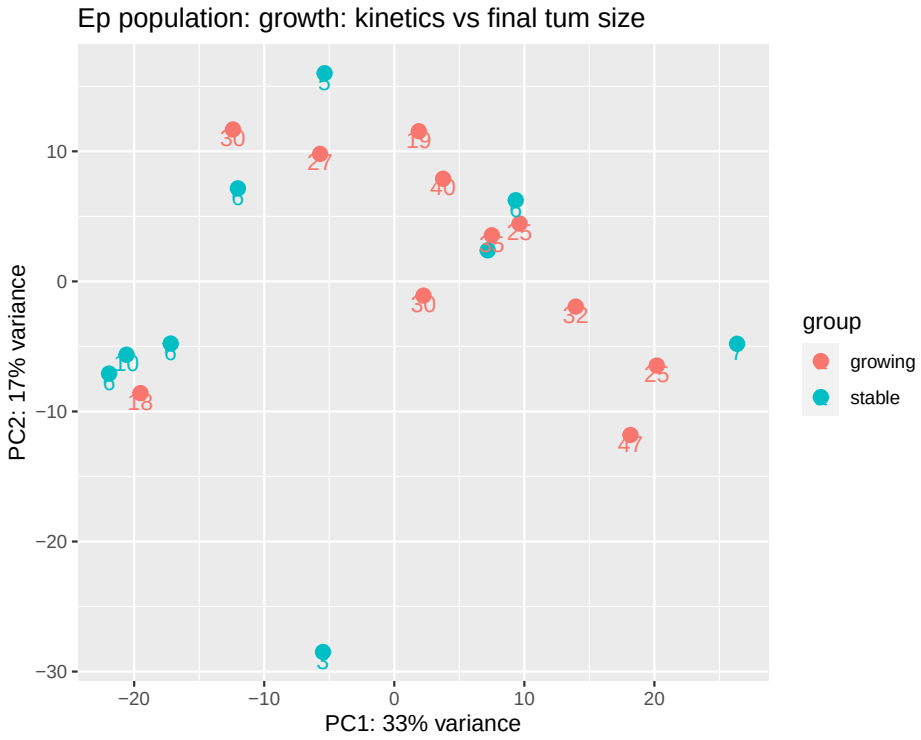
check whether these are needed?

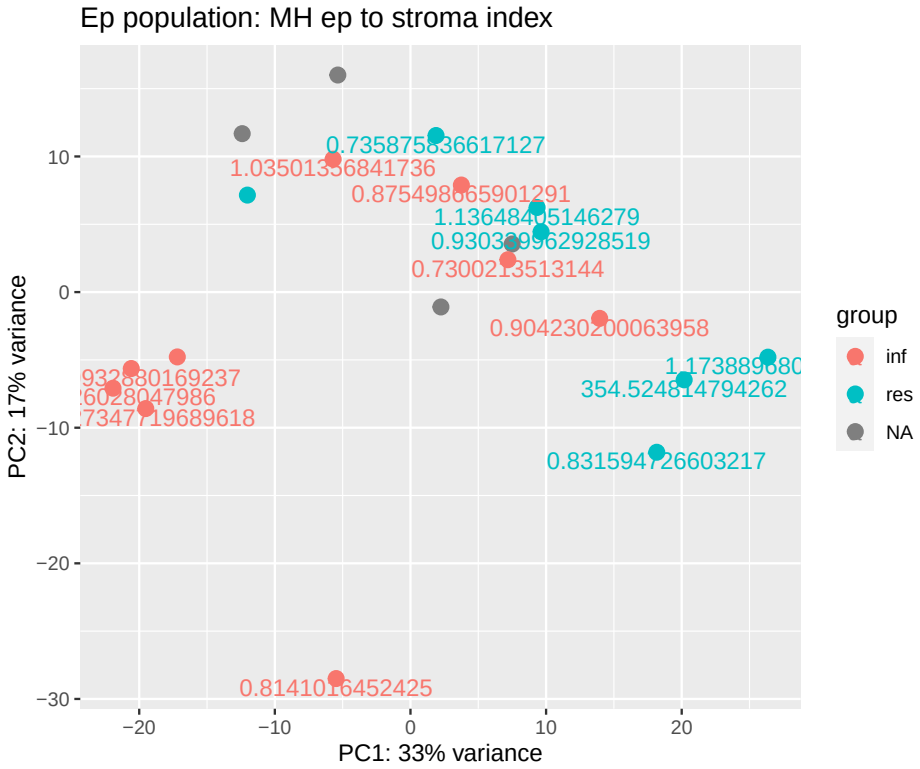
Also check the variance with the different samples

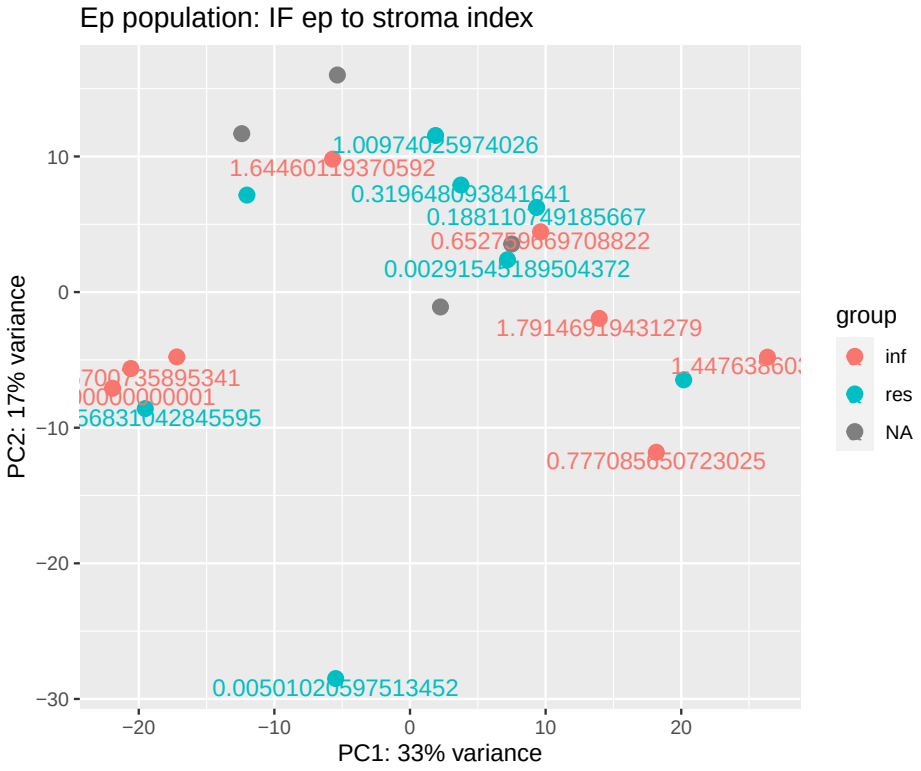

Ep population: treatment
14N_D_Ep
15N_C_Ep
11R_C_Ep
8R_CU_Ep
10R_BL_Ep
12L_D_Ep
14N_C_Ep
11R_D_Ep
17N_D_Ep
7N_A_Ep
3R_C_Ep
16L_C_E
11L_B_Ep
10L_D_Ep
6R_B_Ep
2N_Ep
11N_D_Ep
PC2: 17% variance
PC1: 33% variance
group
LY
PDL1
PDL1+LY
Vehicle

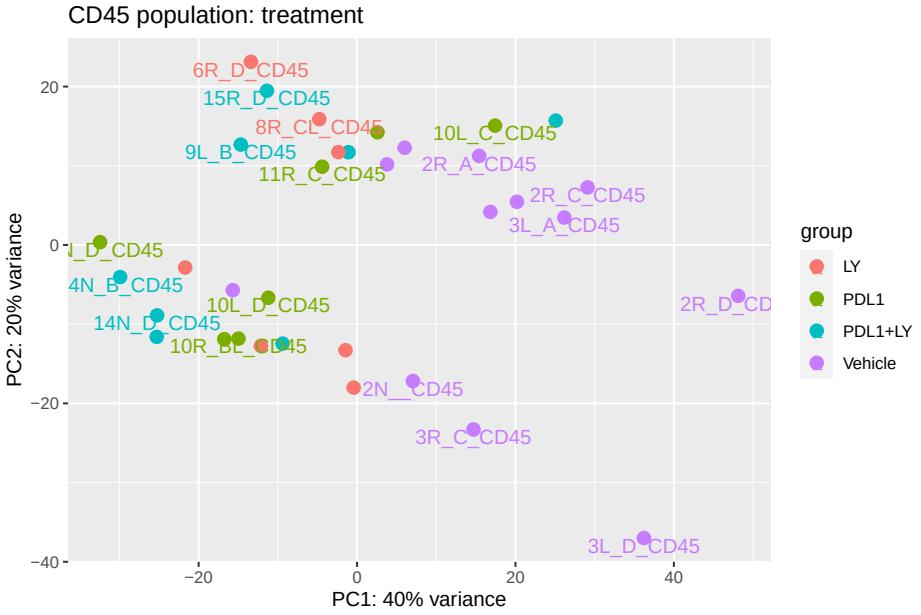

Ep population: treatment + growth
14N_D_Ep
15N_C_Ep
11R_C_Ep
8R_CU_Ep
10R_BL_Ep
12L_D_Ep
14N_C_Ep
11R_D_Ep
17N_D_Ep
7N_A_Ep
3R_C_Ep
16L_C_E
11L_B_Ep
10L_D_Ep
6R_B_Ep
2N_Ep
11N_D_Ep
PC2: 17% variance
PC1: 33% variance
group
LY:growing
LY:stable
PDL1:growing
PDL1:stable
PDL1+LY:growing
PDL1+LY:stable
Vehicle:growing
Vehicle:stable

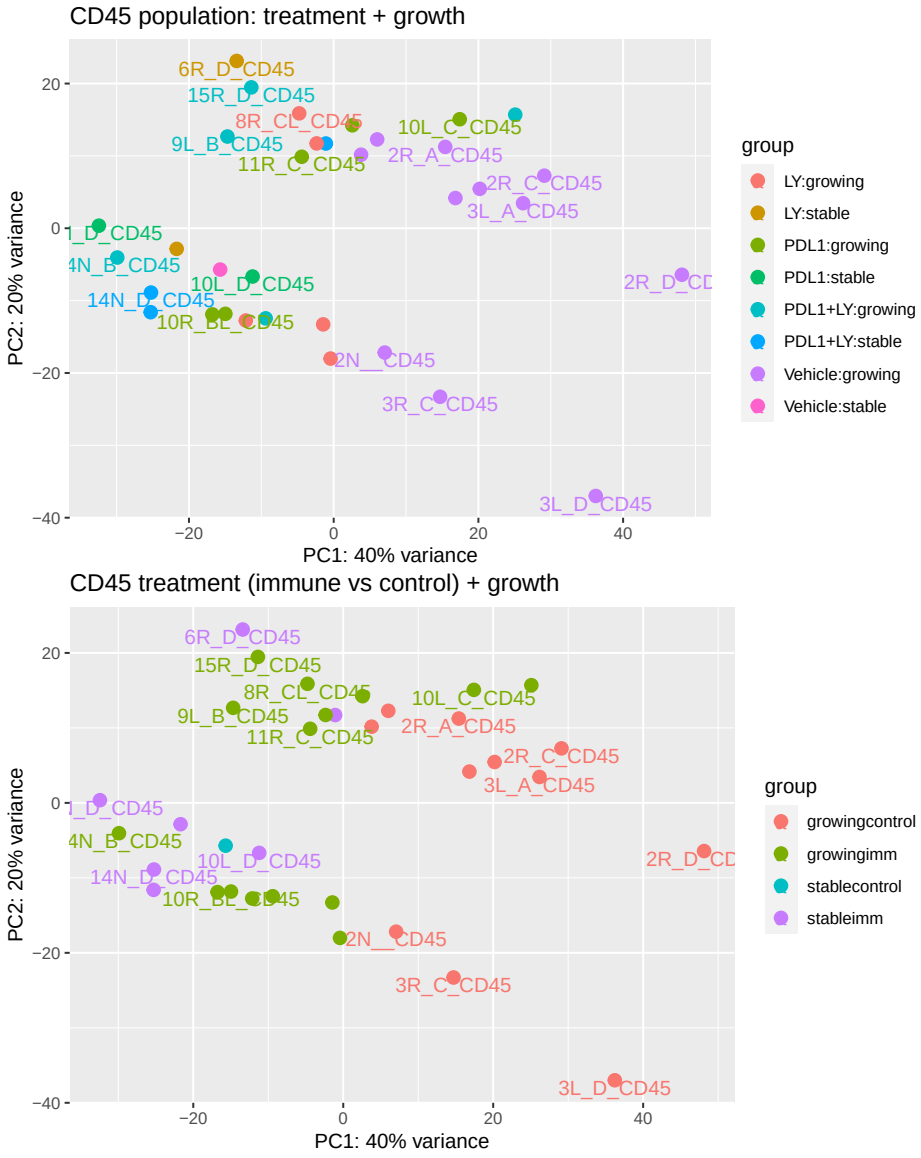

Ep treatment (immune vs control) + growth
14N_D_Ep
15N_C_Ep
11R_C_Ep
8R_CU_Ep
10R_BL_Ep
12L_D_Ep
14N_C_Ep
11R_D_Ep
17N_D_Ep
7N_A_Ep
3R_C_Ep
16L_C_E
11L_B_Ep
6R_B_Ep
2N_Ep
11N_D_Ep
PC2: 17% variance
PC1: 33% variance
group
growingcontrol
growingimm
stablecontrol
stableimm

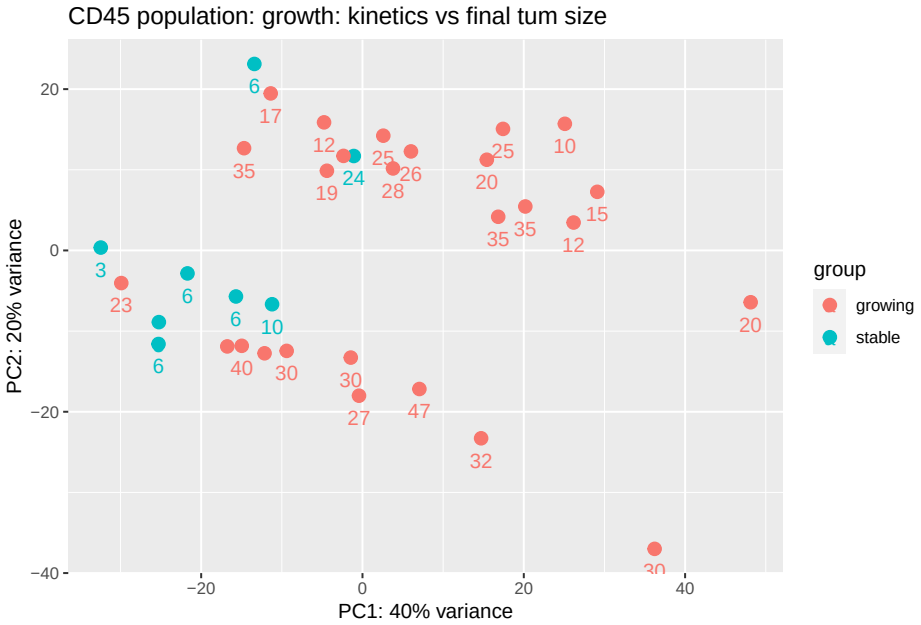

Ep population: growth: kinetics vs final tum size
PC2: 17% variance
PC1: 33% variance
group
growing
stable

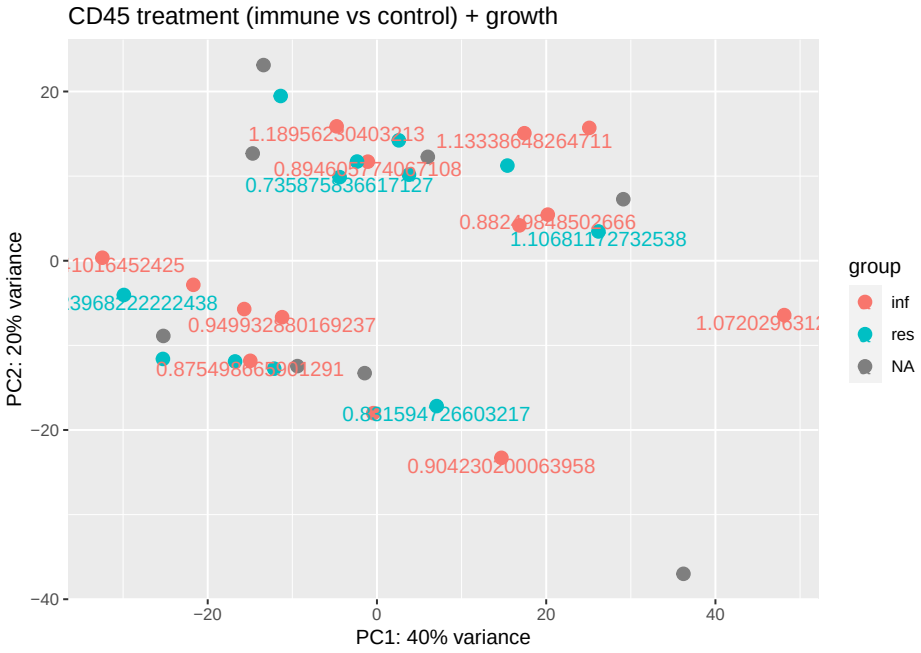

Ep population: MH ep to stroma index
0.7358758636617127
1.03501356841736
0.875498665901291
1.136484705146279
0.9303899962928519
0.7300213513144
0.9042302000639 58
354.524814794262
0.8315947 26603217
0.8141016452425
PC2: 17% variance
PC1: 33% variance
10
0
−10
−20
−30
−20
−10
0
10
20
group
inf
res
NA

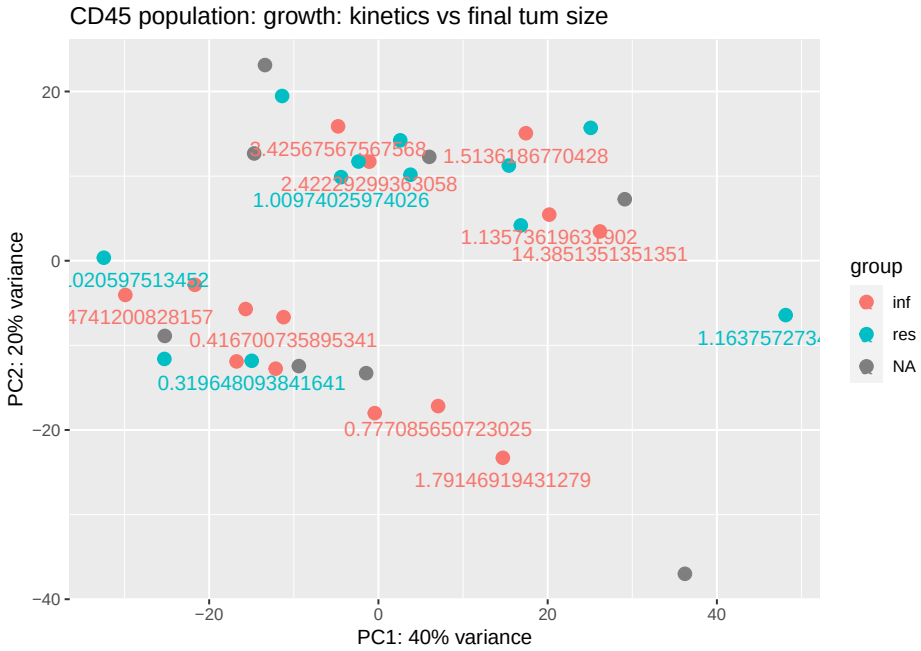

Ep population: IF ep to stroma index
1.00974025974026
1.64460119370592
0.319648093841641
0.188110749185667
0.652789669708822
0.00291545189504372
1.79146919431279
0.770850550723025
0.00501020597513452
group
inf
res
NA
PC2: 17% variance
PC1: 33% variance

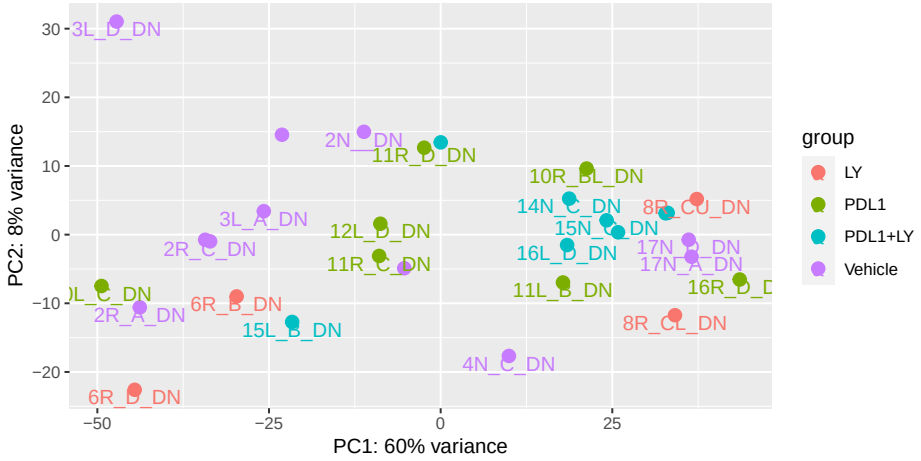

CD45 population: treatment
6R_D_CD45
15R_D_CD45
10L_C_CD45
9L_B_CD45
2R_A_CD45
2R_C_CD45
3L_A_CD45
4N_B_CD45
10L_D_CD45
2R_D_CD
14N_D_CD45
2N__CD45
3R_C_CD45
3L_D_CD45
group
LY
PDL1
PDL1+LY
Vehicle
PC2: 20% variance
PC1: 40% variance

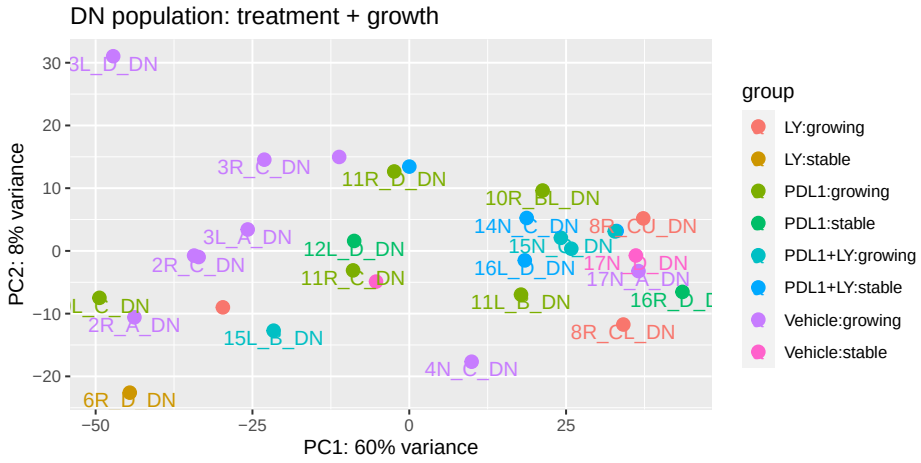

CD45 population: treatment + growth
PC2: 20% variance
PC1: 40% variance
group
LY:growing
LY:stable
PDL1:growing
PDL1:stable
PDL1+LY:growing
PDL1+LY:stable
Vehicle:growing
Vehicle:stable
6R_D_CD45
15R_D_CD45
8R_CL_CD45
10L_C_CD45
9L_B_CD45
11R_C_CD45
2R_A_CD45
2R_C_CD45
3L_A_CD45
10L_D_CD45
14N_D_CD45
2N__CD45
3R_C_CD45
3L_D_CD45
CD45 treatment (immune vs control) + growth
PC2: 20% variance
PC1: 40% variance
group
growingcontrol
growingimm
stablecontrol
stableimm
6R_D_CD45
15R_D_CD45
8R_CL_CD45
10L_C_CD45
9L_B_CD45
11R_C_CD45
2R_A_CD45
2R_C_CD45
3L_A_CD45
10L_D_CD45
14N_D_CD45
2N__CD45
3R_C_CD45
3L_D_CD45

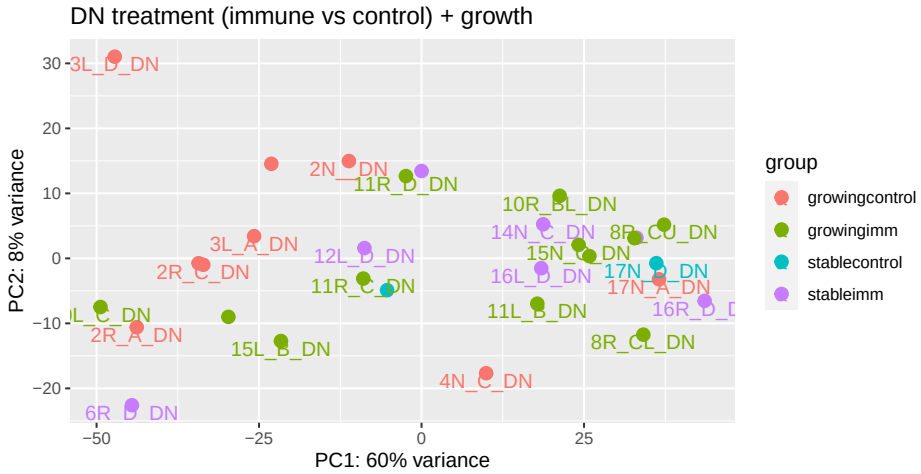

CD45 population: growth: kinetics vs final tum size
PC2: 20% variance
PC1: 40% variance
group
growing
stable

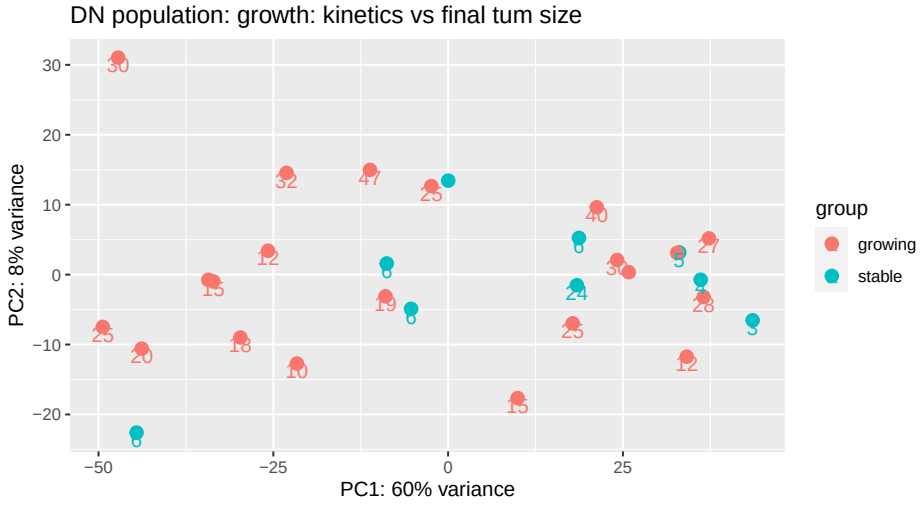

CD45 treatment (immune vs control) + growth
PC2: 20% variance
PC1: 40% variance
group
inf
res
NA

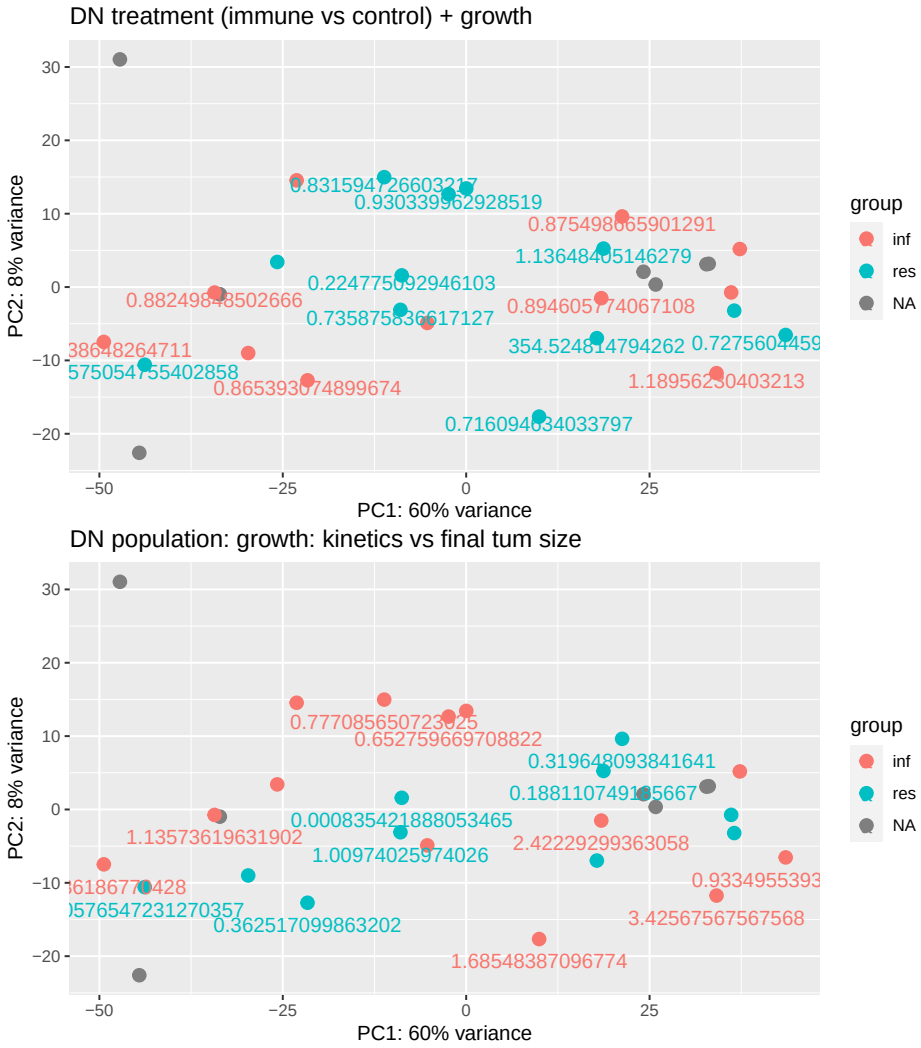

CD45 population: growth: kinetics vs final tum size
PC2: 20% variance
PC1: 40% variance
group
inf
res
NA
3.42567567567568
1.5136186770428
2.42229299363058
1.00974025974026
1.13573619631902
14.3851351351351
0.416700735895341
0.319648093841641
0.777085650723025
1.79146919431279


DN population: treatment
PC2: 8% variance
PC1: 60% variance
group
LY
PDL1
PDL1+LY
Vehicle
3L_D_DN
2N_C_DN
11R_D_DN
10R_B_L_DN
14N_C_DN
8R_C_U_DN
3L_A_DN
12L_D_DN
2R_C_DN
11R_C_DN
16L_D_DN
17N_A_DN
11L_B_DN
2R_A_DN
6R_B_DN
15L_B_DN
8R_C_L_DN
4N_C_DN
6R_D_DN


DN population: treatment + growth
PC2: 8% variance
PC1: 60% variance
group
LY:growing
LY:stable
PDL1:growing
PDL1:stable
PDL1+LY:growing
PDL1+LY:stable
Vehicle:growing
Vehicle:stable


DN treatment (immune vs control) + growth
PC2: 8% variance
PC1: 60% variance
group
growingcontrol
growingimm
stablecontrol
stableimm


DN population: growth: kinetics vs final tum size
PC2: 8% variance
PC1: 60% variance
group
growing
stable



7.3.1 Results : differences in the epithelial fraction

Save to file the differences in the epithelial fractions

```
##
## growingcontrol     growingimm  stablecontrol      stableimm
##              2              9              2              7
```

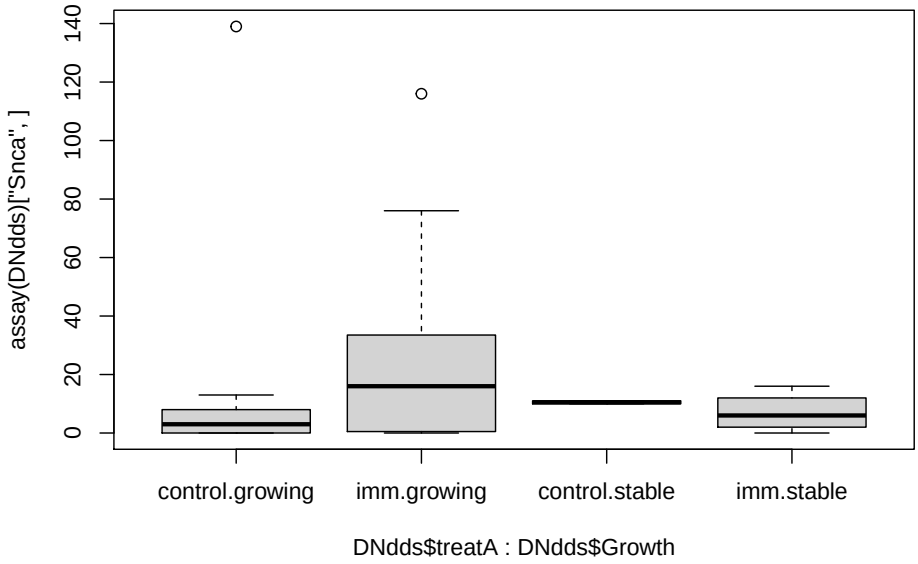
7.4 GSEA

- may need to convert gene names into HUGO symbols

Prepare for GSEA: This takes approximately 15 minutes

7.5 CD45 comparisons

7.6 DN comparisons



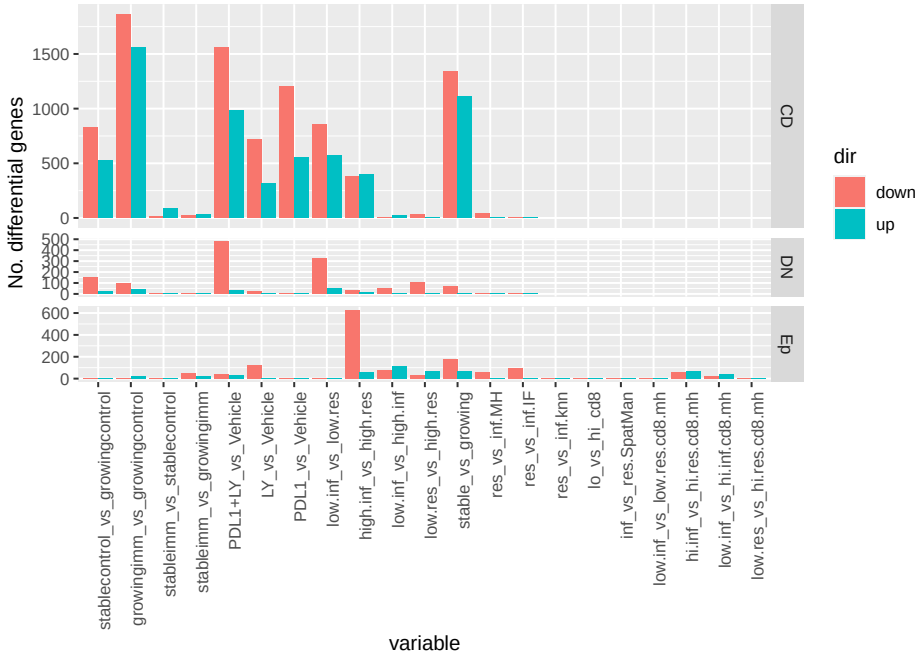
Chapter 8

DESeq analysis

8.0.1 Summary plots

Quick plot of the differences in number of differential genes:

```
##   Frac  dir                        variable value
## 1   Ep down stablecontrol_vs_growingcontrol     1
## 2   Ep   up stablecontrol_vs_growingcontrol     1
## 3   CD down stablecontrol_vs_growingcontrol   830
## 4   CD   up stablecontrol_vs_growingcontrol   524
## 5   DN down stablecontrol_vs_growingcontrol   152
## 6   DN   up stablecontrol_vs_growingcontrol    22
```




```
##   Frac  dir                          variable value
## 1 <NA> <NA> stablecontrol_vs_growingcontrol     2
## 2 <NA> <NA> stablecontrol_vs_growingcontrol     2
## 3 <NA> <NA> stablecontrol_vs_growingcontrol     1
## 4 <NA> <NA> stablecontrol_vs_growingcontrol    11
## 5 <NA> <NA> stablecontrol_vs_growingcontrol     2
## 6 <NA> <NA> stablecontrol_vs_growingcontrol     9
```

Any commonly changed genes in any of the lists? Plot lists: * Pearson correlation of all the samples * Unlist all, and pick out genes which are impacted in at least 1 comparison

CD45 comparisons

DN samples:

8.1 Focus mainly on growing vs stable:

Check: Is the difference between stable and growing just MHC expression?

8.1.1 comapre DN samples: PAM50 Lum vs Basal/Normal?