```r
library(ggplot2)
library(ggpubr)
```

```
## Loading required package: magrittr
```

```r
library(ggpointdensity)
theme_set(theme_bw())

# ------------------------------------------------------------------------
# load results of deconvolution
# ------------------------------------------------------------------------

load("deconv_results/deconv_expr_COAD.RData")
ls()
```

```
## [1] "deconv_expr_COAD"
```

```r
load("deconv_results/deconv_methy_COAD.RData")
ls()
```

```
## [1] "deconv_expr_COAD" "rho_COAD"
```

```r
load("deconv_results/barcode/subsample_coad.RData")
ls()
```

```
## [1] "deconv_expr_COAD" "rho_COAD"         "subsample.coad"
```

```r
dim(deconv_expr_COAD)
```

```
## [1] 175   7
```

```r
dim(rho_COAD)
```

```
## [1] 175   7   5
```

```r
dim(subsample.coad)
```

```
## [1] 175   3
```

```r
deconv_expr_COAD[1:2,]
```

```
##            CD4T       CD8T   Monocyte         B         NK  Neutrophil
## 5656 0.10297062 0.02254813 0.03457699 0.0591461 0.02676888 0.008814001
## 6781 0.06966723 0.14954229 0.12023085 0.0921302 0.06659492 0.057396469
##            Treg
## 5656 0.01517529
## 6781 0.02443803
```

```r
rho_COAD[1,,]
```

```
##                   EMeth        svr         ls          rls            qp
## CD4T        4.522374e-02 0.08870032 0.00000000 0.0000000000  2.700000e-01
## CD8T        9.513481e-20 0.05843881 0.14527590 0.1433246771 -2.420395e-16
## Monocyte    6.179749e-02 0.05511302 0.04652121 0.0493472750 -9.639956e-17
## B           2.972052e-02 0.02804754 0.03134310 0.0336701679  0.000000e+00
## NK         -1.703507e-19 0.00707734 0.00000000 0.0004295355  1.299862e-16
## Neutrophil  5.421388e-02 0.01964418 0.04685979 0.0432283445 -1.481225e-16
## Treg        7.904436e-02 0.01297878 0.00000000 0.0000000000 -8.181769e-17
```

```r
subsample.coad[1:2,]
```

```
##   num patient_id      barcode
## 1   1       2671 TCGA-A6-2671
## 2   2       2675 TCGA-A6-2675
```

```r
dimnames(rho_COAD)
```

```
## [[1]]
##   [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"  "12"
##  [13] "13"  "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"  "23"  "24"
##  [25] "25"  "26"  "27"  "28"  "29"  "30"  "31"  "32"  "33"  "34"  "35"  "36"
##  [37] "37"  "38"  "39"  "40"  "41"  "42"  "43"  "44"  "45"  "46"  "47"  "48"
##  [49] "49"  "50"  "51"  "52"  "53"  "54"  "55"  "56"  "57"  "58"  "59"  "60"
##  [61] "61"  "62"  "63"  "64"  "65"  "66"  "67"  "68"  "69"  "70"  "71"  "72"
##  [73] "73"  "74"  "75"  "76"  "77"  "78"  "79"  "80"  "81"  "82"  "83"  "84"
##  [85] "85"  "86"  "87"  "88"  "89"  "90"  "91"  "92"  "93"  "94"  "95"  "96"
##  [97] "97"  "98"  "99"  "100" "101" "102" "103" "104" "105" "106" "107" "108"
## [109] "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" "119" "120"
## [121] "121" "122" "123" "124" "125" "126" "127" "128" "129" "130" "131" "132"
## [133] "133" "134" "135" "136" "137" "138" "139" "140" "141" "142" "143" "144"
## [145] "145" "146" "147" "148" "149" "150" "151" "152" "153" "154" "155" "156"
## [157] "157" "158" "159" "160" "161" "162" "163" "164" "165" "166" "167" "168"
## [169] "169" "170" "171" "172" "173" "174" "175"
##
## [[2]]
## [1] "CD4T"       "CD8T"       "Monocyte"   "B"          "NK"
## [6] "Neutrophil" "Treg"
##
## [[3]]
## [1] "EMeth" "svr"   "ls"    "rls"   "qp"
```

```r
table(rownames(deconv_expr_COAD) == subsample.coad$patient_id)
```

```
##
## FALSE  TRUE
##   172     3
```

```r
cor(deconv_expr_COAD[,"B"], rho_COAD[,"B","EMeth"])
```

```
## [1] 0.7301608
```

```r
# ------------------------------------------------------------------------
# load results of clinical information
# ------------------------------------------------------------------------

sc = readRDS("clinical_data/COAD_somatic_clinic.rds")
dim(sc)
```

```
## [1] 384  38
```

```r
names(sc)
```

```
##  [1] "barcode"
##  [2] "SMASH_S_hg38"
##  [3] "SMASH_oE_hg38"
##  [4] "SMASH_wE_hg38"
##  [5] "SMASH_oNE_hg38"
```

```
##  [6] "SMASH_wNE_hg38"
##  [7] "AscatPurity"
##  [8] "AscatPloidy"
##  [9] "raw_MB_hg38_SNV"
## [10] "raw_MB_hg38_INDEL"
## [11] "IDH_CNV_status_hg38"
## [12] "tCN_burden"
## [13] "tCN_burden_ap"
## [14] "tumor_type.x"
## [15] "num_clonal"
## [16] "num_subclonal"
## [17] "prop_clonal"
## [18] "gender"
## [19] "age"
## [20] "tumor_type.y"
## [21] "initial_pathologic_diagnosis_method"
## [22] "histological_type"
## [23] "pathologic_stage"
## [24] "gleason_score"
## [25] "psa_level"
## [26] "hpv_status_by_ish_testing"
## [27] "hpv_status_by_p16_testing"
## [28] "prior_glioma"
## [29] "history_of_neoadjuvant_treatment"
## [30] "breast_carcinoma_estrogen_receptor_status"
## [31] "breast_carcinoma_progesterone_receptor_status"
## [32] "neoplasm_histologic_grade"
## [33] "yr_of_tobacco_smoking_onset"
## [34] "pathologic_T"
## [35] "stage"
## [36] "Time"
## [37] "Delta"
## [38] "radiation_therapy"
```

```
sc = sc[,-(23:32)]
dim(sc)
```

```
## [1] 384  28
```

```
sc[1:2,]
```

```
##        barcode SMASH_S_hg38 SMASH_oE_hg38 SMASH_wE_hg38 SMASH_oNE_hg38
## 1 TCGA-3L-AA1B            2          0.69     0.6765692      0.2783978
## 2 TCGA-4N-A93T            3          0.95     1.0292827      0.1656512
##   SMASH_wNE_hg38 AscatPurity AscatPloidy raw_MB_hg38_SNV raw_MB_hg38_INDEL
## 1      0.1958878        0.41    3.712254              94                 2
## 2      0.1803782        0.82    1.847516              69                 4
##   IDH_CNV_status_hg38 tCN_burden tCN_burden_ap tumor_type.x num_clonal
## 1       IDH wild type  1.7672127     0.8270333         COAD         88
## 2       IDH wild type  0.3889078     0.4157566         COAD         56
##   num_subclonal prop_clonal gender age tumor_type.y
## 1             6   0.9361702 FEMALE  61        Colon
## 2            13   0.8115942   MALE  67        Colon
##   initial_pathologic_diagnosis_method    histological_type
## 1                                <NA> Colon Adenocarcinoma
## 2                                <NA> Colon Adenocarcinoma
```

```
##   yr_of_tobacco_smoking_onset pathologic_T stage   Time Delta radiation_therapy
## 1                          NA          T2     I 475.01     0                NO
## 2                          NA         T4a   III 146.01     0                NO
```

```
ff0    = "clinical_data/patient_coad_M_info_hyperMeth.txt"
emInfo = read.table(ff0, sep = "\t", header = TRUE, as.is = TRUE)
dim(emInfo)
```

```
## [1] 213  28
```

```
emInfo[1, ]
```

```
##    race gender              ethnicity bcr_patient_barcode patient_id
## 1 WHITE   MALE NOT HISPANIC OR LATINO        TCGA-A6-2671       2671
##   tissue_source_site birth_days_to last_contact_days_to death_days_to
## 1                 A6        -31329                  648          1331
##   vital_status tumor_status ajcc_pathologic_tumor_stage
## 1         Dead   WITH TUMOR                    Stage IV
##   age_at_initial_pathologic_diagnosis     noHM_PC1    noHM_PC2   noHM_PC3
## 1                                  85 -0.009249593 0.004130469 0.01562243
##      noHM_PC4          methylation_barcode
## 1 0.003222672 TCGA-A6-2671-01A-01D-1407-05
##                                                                 methylation_file
## 1 jhu-usc.edu_COAD.HumanMethylation450.1.lvl-3.TCGA-A6-2671-01A-01D-1407-05.txt
##              platform                                                      array
## 1 HumanMethylation450 RARER_p_TCGA_MixedRedos_N_GenomeWideSNP_6_D05_747712
##   segment_count abs_call abs_purity abs_ploidy abs_doublings hyperMeth
## 1           182   called       0.73       3.57             1         0
##   CIMP.Status
## 1    Negative
```

```
table(emInfo$bcr_patient_barcode %in% sc$barcode)
```

```
##
## FALSE  TRUE
##     9   204
```

```
names(emInfo)[which(names(emInfo)=="bcr_patient_barcode")] = "barcode"

emInfo = merge(emInfo, sc, by="barcode")
dim(emInfo)
```

```
## [1] 204  55
```

```
table(emInfo$vital_status, emInfo$Delta)
```

```
##
##           0   1
##   Alive 152   0
##   Dead    0  52
```

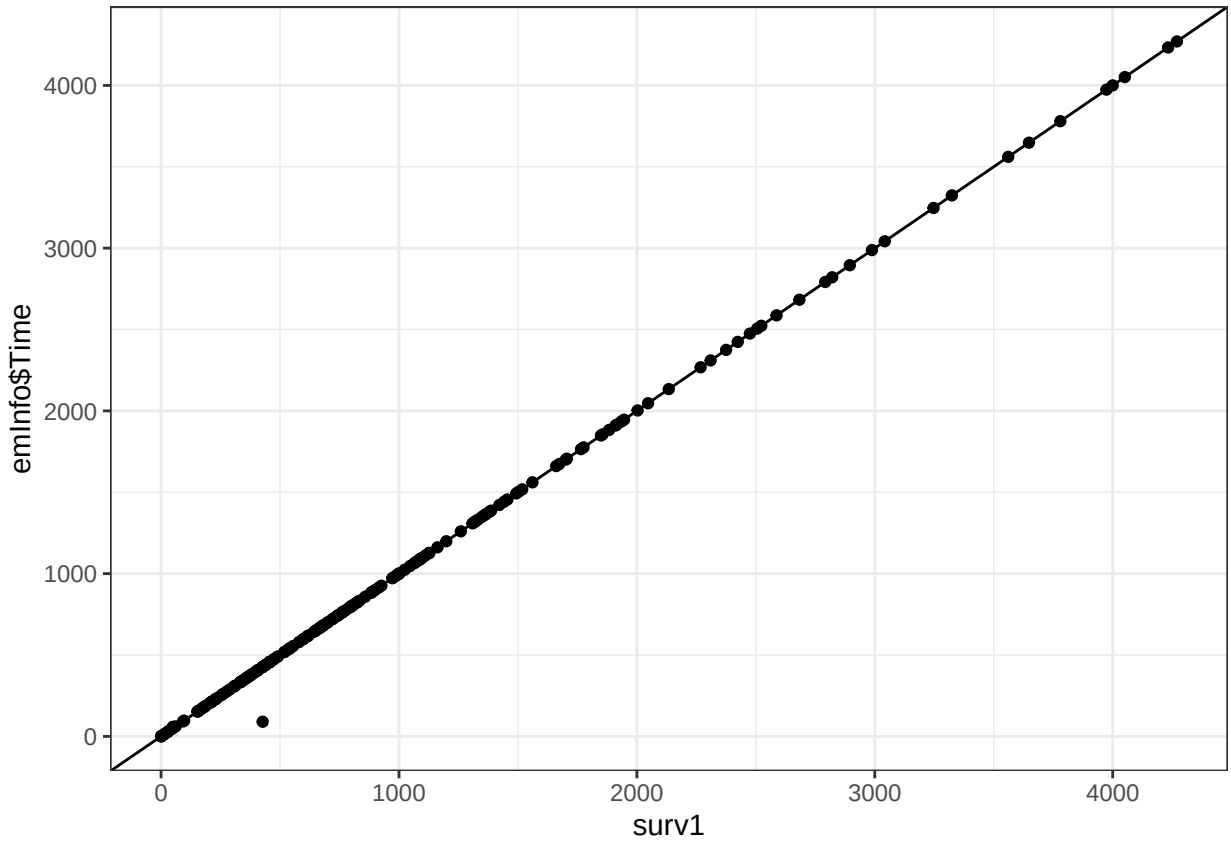
```
surv1 = pmax(emInfo$last_contact_days_to, emInfo$death_days_to, na.rm=TRUE)
summary(surv1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   391.2   753.5  1029.3  1350.8  4270.0
```

```
summary(emInfo$Time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.01  382.26  753.51 1027.71 1350.76 4270.01
```

```
emInfo$Time = emInfo$Time - 0.01
emInfo$Time[which(emInfo$Time == 0)] = 1

# confirm the survival time are consistent between the two studies
ggplot(emInfo, aes(x=surv1, y=emInfo$Time)) + geom_point() +
  geom_abline(intercept=0, slope=1)
```




```
# ------------------------------------------------------------------------
# merage clinical information and cell type composition estimates
# ------------------------------------------------------------------------

colnames(deconv_expr_COAD) = paste0(colnames(deconv_expr_COAD), ".E")
ct = cbind(subsample.coad, deconv_expr_COAD, rho_COAD[,,"EMeth"])
dim(ct)
```

```
## [1] 175  17
```

```
ct[1:2,]
```

```
##      num patient_id      barcode     CD4T.E     CD8T.E Monocyte.E       B.E
## 5656   1       2671 TCGA-A6-2671 0.10297062 0.02254813 0.03457699 0.0591461
## 6781   2       2675 TCGA-A6-2675 0.06966723 0.14954229 0.12023085 0.0921302
##            NK.E Neutrophil.E     Treg.E       CD4T         CD8T   Monocyte
## 5656 0.02676888  0.008814001 0.01517529 0.04522374 9.513481e-20 0.06179749
## 6781 0.06659492  0.057396469 0.02443803 0.17210686 9.463496e-02 0.06969726
##               B            NK Neutrophil       Treg
```

```
## 5656 0.02972052 -1.703507e-19 0.05421388 0.07904436
## 6781 0.07070684  3.419027e-02 0.03432683 0.10433699
```

```
table(rownames(ct) == ct$patient_id)
```

```
##
## FALSE  TRUE
##   172     3
```

```
table(ct$barcode %in% emInfo$barcode)
```

```
##
## FALSE  TRUE
##     5   170
```

```
ct = merge(ct, emInfo, by="barcode")
dim(ct)
```

```
## [1] 170  71
```

```
ct[1:2,]
```

```
##        barcode num patient_id.x     CD4T.E     CD8T.E Monocyte.E       B.E
## 1 TCGA-A6-2671   1         2671 0.10297062 0.02254813 0.03457699 0.0591461
## 2 TCGA-A6-2675   2         2675 0.06966723 0.14954229 0.12023085 0.0921302
##         NK.E Neutrophil.E     Treg.E       CD4T         CD8T   Monocyte
## 1 0.02676888  0.008814001 0.01517529 0.04522374 9.513481e-20 0.06179749
## 2 0.06659492  0.057396469 0.02443803 0.17210686 9.463496e-02 0.06969726
##            B            NK Neutrophil       Treg  race gender.x
## 1 0.02972052 -1.703507e-19 0.05421388 0.07904436 WHITE     MALE
## 2 0.07070684  3.419027e-02 0.03432683 0.10433699 WHITE     MALE
##                ethnicity patient_id.y tissue_source_site birth_days_to
## 1 NOT HISPANIC OR LATINO         2671                 A6        -31329
## 2 NOT HISPANIC OR LATINO         2675                 A6        -28813
##   last_contact_days_to death_days_to vital_status tumor_status
## 1                  648          1331         Dead   WITH TUMOR
## 2                 1321            NA        Alive   TUMOR FREE
##   ajcc_pathologic_tumor_stage age_at_initial_pathologic_diagnosis     noHM_PC1
## 1                    Stage IV                                  85 -0.009249593
## 2                   Stage IIA                                  78  0.098222522
##      noHM_PC2   noHM_PC3     noHM_PC4          methylation_barcode
## 1 0.004130469 0.01562243 0.0032226720 TCGA-A6-2671-01A-01D-1407-05
## 2 0.023360098 0.04020544 0.0009877077 TCGA-A6-2675-01A-02D-1721-05
##                                                        methylation_file
## 1 jhu-usc.edu_COAD.HumanMethylation450.1.lvl-3.TCGA-A6-2671-01A-01D-1407-05.txt
## 2 jhu-usc.edu_COAD.HumanMethylation450.3.lvl-3.TCGA-A6-2675-01A-02D-1721-05.txt
##             platform                                                  array
## 1 HumanMethylation450 RARER_p_TCGA_MixedRedos_N_GenomeWideSNP_6_D05_747712
## 2 HumanMethylation450   GRIPS_p_TCGA_b116_SNP_N_GenomeWideSNP_6_C07_781418
##   segment_count abs_call abs_purity abs_ploidy abs_doublings hyperMeth
## 1           182   called       0.73       3.57             1         0
## 2           230   called       0.42       3.37             1         0
##   CIMP.Status SMASH_S_hg38 SMASH_oE_hg38 SMASH_wE_hg38 SMASH_oNE_hg38
## 1    Negative            3          1.06     0.9274599      0.5531430
## 2    Negative            2          0.68     0.7213201      0.6600493
##   SMASH_wNE_hg38 AscatPurity AscatPloidy raw_MB_hg38_SNV raw_MB_hg38_INDEL
## 1      0.4846527        0.68    3.609337              49                 7
```

```
## 2      0.6585205        0.49    3.351309              72                 2
##   IDH_CNV_status_hg38 tCN_burden tCN_burden_ap tumor_type.x num_clonal
## 1       IDH wild type   2.228662     1.0712969         COAD         39
## 2       IDH wild type   1.572784     0.8041944         COAD         46
##   num_subclonal prop_clonal gender.y age tumor_type.y
## 1            10   0.7959184     MALE  85        Colon
## 2            26   0.6388889     MALE  78        Colon
##   initial_pathologic_diagnosis_method    histological_type
## 1                                <NA> Colon Adenocarcinoma
## 2                                <NA> Colon Adenocarcinoma
##   yr_of_tobacco_smoking_onset pathologic_T stage Time Delta radiation_therapy
## 1                          NA           T3    IV 1331     1                NO
## 2                          NA           T3    II 1321     0                NO
```

```
names(ct)
```

```
##  [1] "barcode"                                "num"
##  [3] "patient_id.x"                           "CD4T.E"
##  [5] "CD8T.E"                                 "Monocyte.E"
##  [7] "B.E"                                    "NK.E"
##  [9] "Neutrophil.E"                           "Treg.E"
## [11] "CD4T"                                   "CD8T"
## [13] "Monocyte"                               "B"
## [15] "NK"                                     "Neutrophil"
## [17] "Treg"                                   "race"
## [19] "gender.x"                               "ethnicity"
## [21] "patient_id.y"                           "tissue_source_site"
## [23] "birth_days_to"                          "last_contact_days_to"
## [25] "death_days_to"                          "vital_status"
## [27] "tumor_status"                           "ajcc_pathologic_tumor_stage"
## [29] "age_at_initial_pathologic_diagnosis"    "noHM_PC1"
## [31] "noHM_PC2"                               "noHM_PC3"
## [33] "noHM_PC4"                               "methylation_barcode"
## [35] "methylation_file"                       "platform"
## [37] "array"                                  "segment_count"
## [39] "abs_call"                               "abs_purity"
## [41] "abs_ploidy"                             "abs_doublings"
## [43] "hyperMeth"                              "CIMP.Status"
## [45] "SMASH_S_hg38"                           "SMASH_oE_hg38"
## [47] "SMASH_wE_hg38"                          "SMASH_oNE_hg38"
## [49] "SMASH_wNE_hg38"                         "AscatPurity"
## [51] "AscatPloidy"                            "raw_MB_hg38_SNV"
## [53] "raw_MB_hg38_INDEL"                      "IDH_CNV_status_hg38"
## [55] "tCN_burden"                             "tCN_burden_ap"
## [57] "tumor_type.x"                           "num_clonal"
## [59] "num_subclonal"                          "prop_clonal"
## [61] "gender.y"                               "age"
## [63] "tumor_type.y"                           "initial_pathologic_diagnosis_method"
## [65] "histological_type"                      "yr_of_tobacco_smoking_onset"
## [67] "pathologic_T"                           "stage"
## [69] "Time"                                   "Delta"
## [71] "radiation_therapy"
```

```
purity.E = 1 - rowSums(ct[,4:10])
purity.M = 1 - rowSums(ct[,11:17])
```

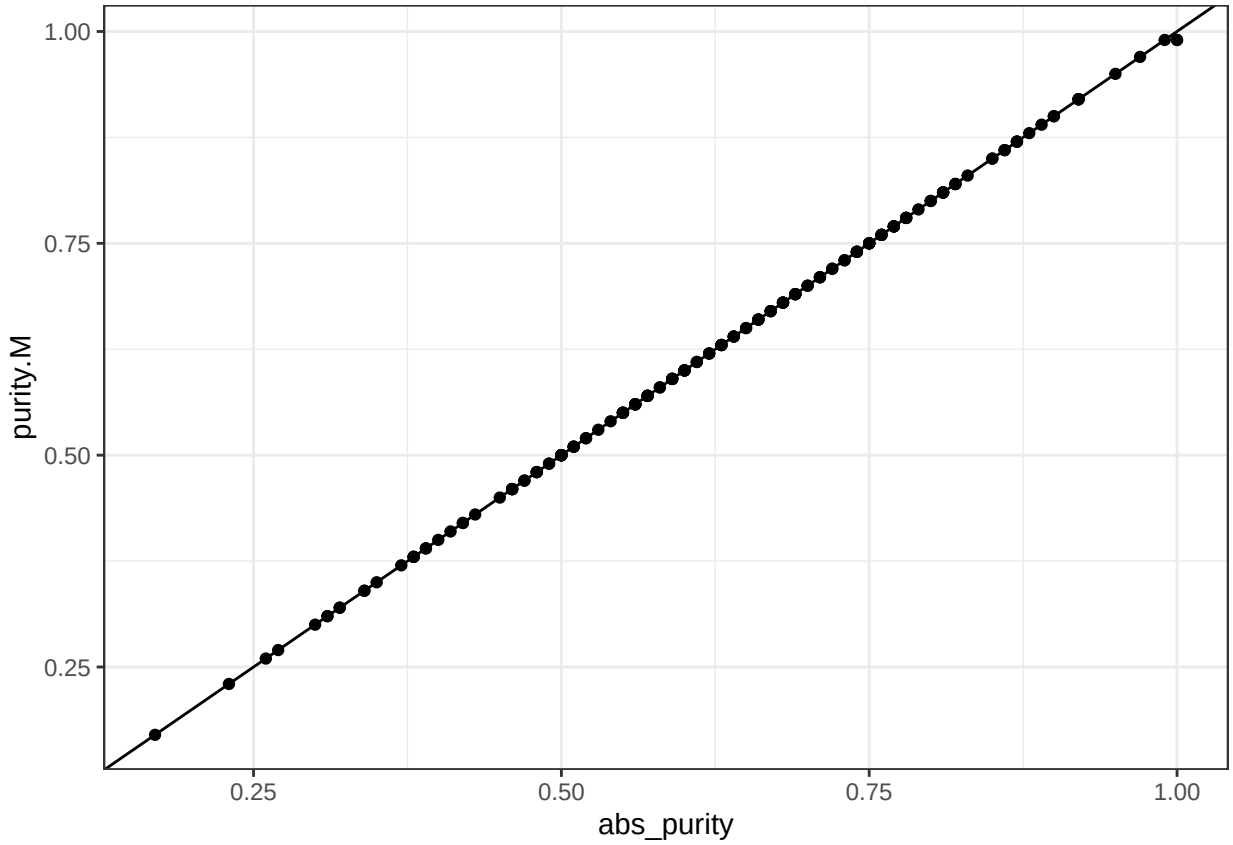
```r
cor(purity.E, purity.M)
```

```
## [1] 0.9999816
```

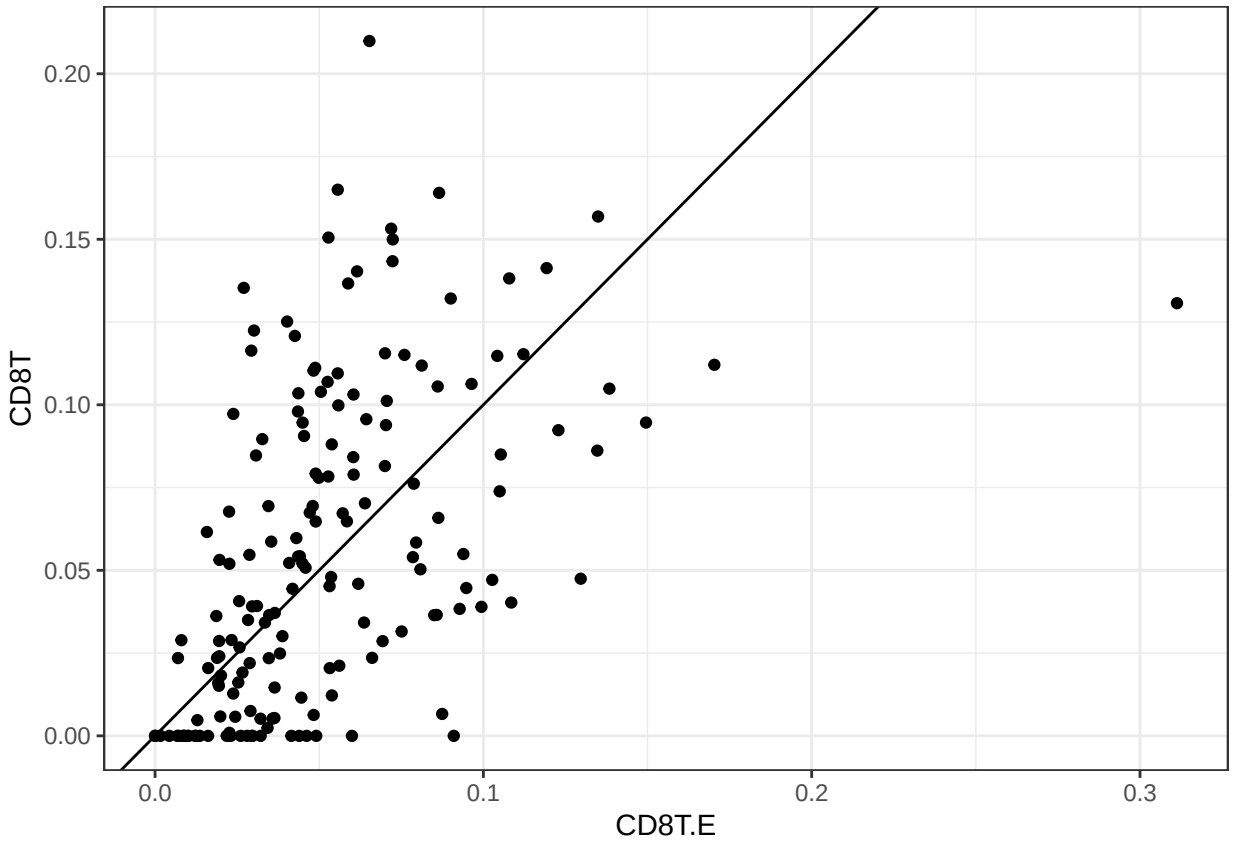
```r
cor(ct$abs_purity, purity.M)
```

```
## [1] 0.9999816
```

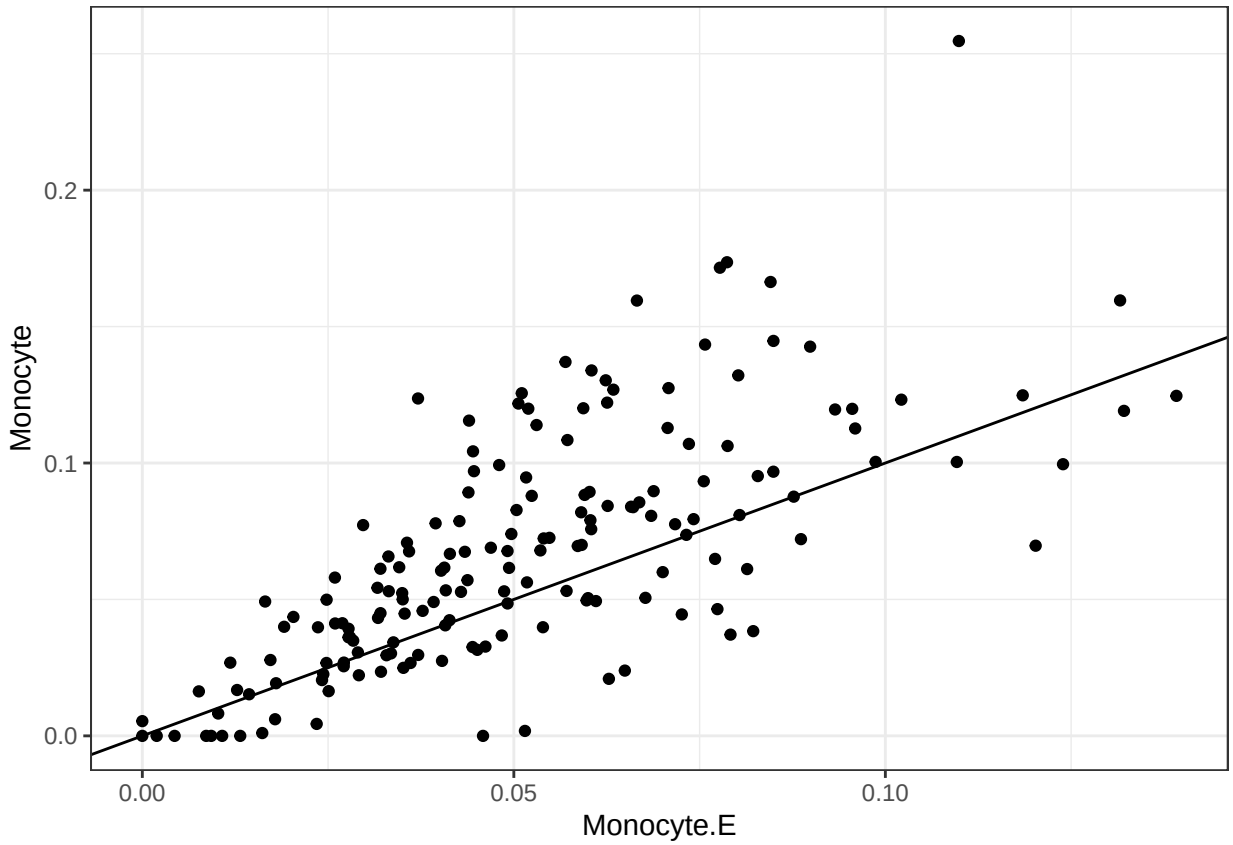
```r
ggplot(ct, aes(x=abs_purity, y=purity.M)) + geom_point() +
  geom_abline(intercept=0, slope=1)
```




```r
ggplot(ct, aes(x=CD8T.E, y=CD8T)) + geom_point() +
  geom_abline(intercept=0, slope=1)
```

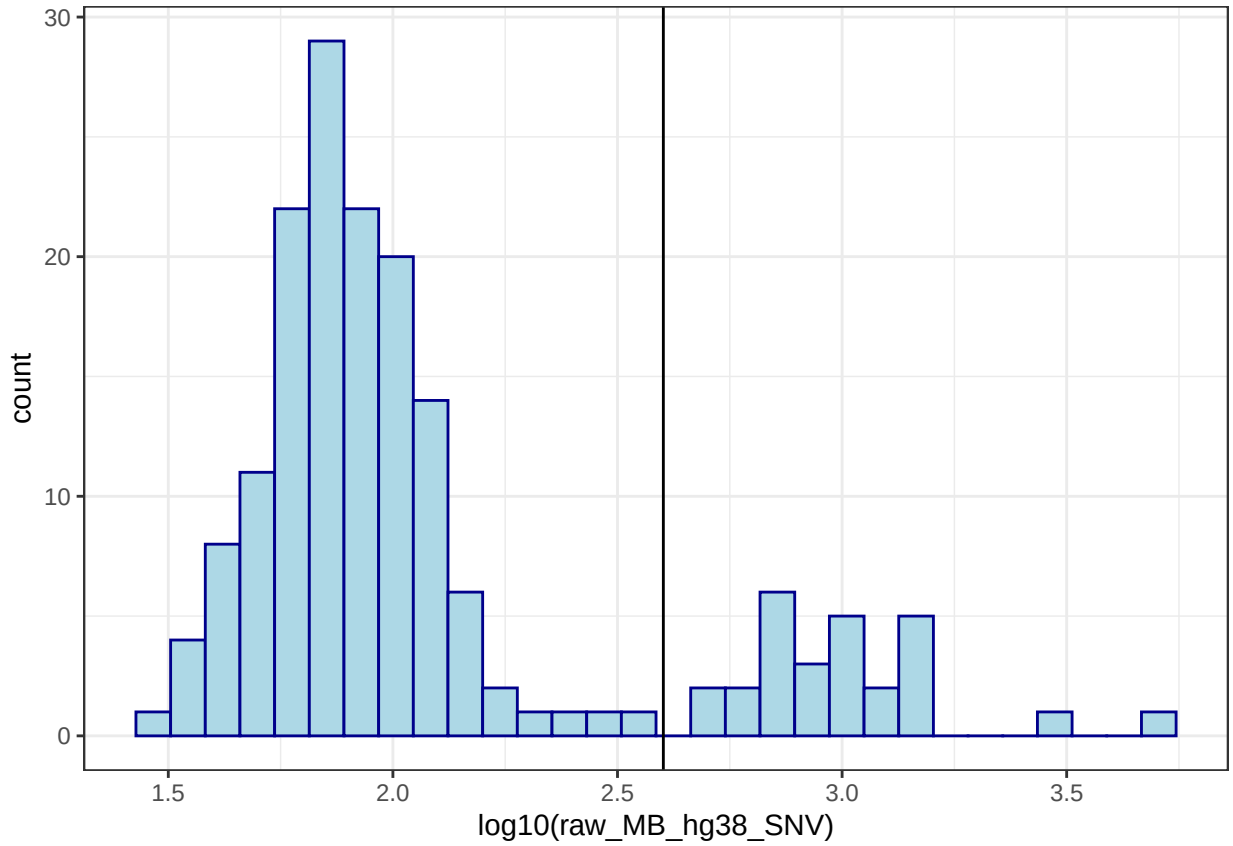



```
ggplot(ct, aes(x=Monocyte.E, y=Monocyte)) + geom_point() +
  geom_abline(intercept=0, slope=1)
```

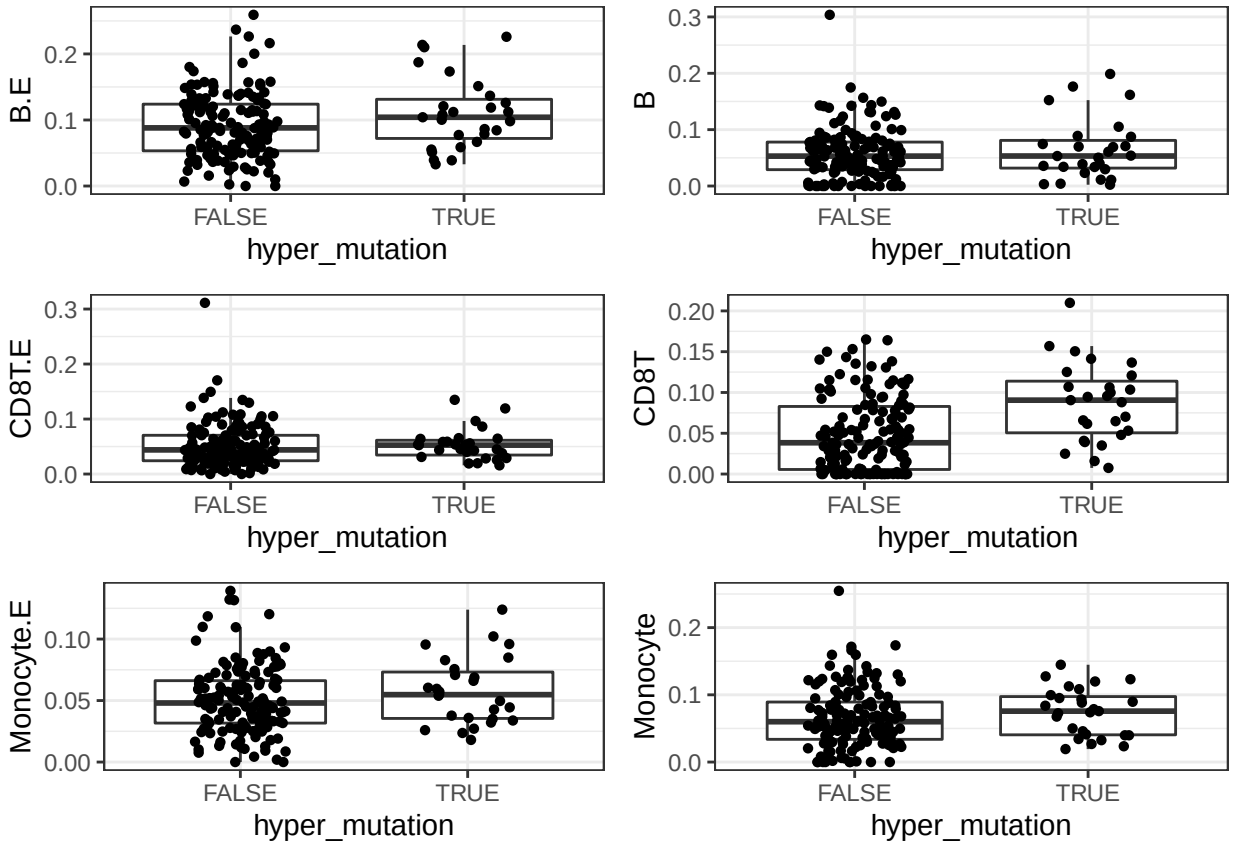



```
ggplot(ct, aes(x=log10(raw_MB_hg38_SNV)))+
  geom_histogram(color="darkblue", fill="lightblue")  +
  geom_vline(xintercept = log10(400))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ct$hyper_mutation = ct[,"raw_MB_hg38_SNV"] > 400

g1 = ggplot(ct, aes(x=hyper_mutation, y=B.E)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(shape=16, position=position_jitter(0.2))

g2 = ggplot(ct, aes(x=hyper_mutation, y=B)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(shape=16, position=position_jitter(0.2))

g3 = ggplot(ct, aes(x=hyper_mutation, y=CD8T.E)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(shape=16, position=position_jitter(0.2))

g4 = ggplot(ct, aes(x=hyper_mutation, y=CD8T)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(shape=16, position=position_jitter(0.2))

g5 = ggplot(ct, aes(x=hyper_mutation, y=Monocyte.E)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(shape=16, position=position_jitter(0.2))

g6 = ggplot(ct, aes(x=hyper_mutation, y=Monocyte)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(shape=16, position=position_jitter(0.2))

ggarrange(g1, g2, g3, g4, g5, g6, nrow = 3, ncol = 2)
```


B.E
0.0
0.1
0.2
FALSE
TRUE
hyper_mutation
B
0.0
0.1
0.2
0.3
FALSE
TRUE
hyper_mutation
CD8T.E
0.0
0.1
0.2
0.3
FALSE
TRUE
hyper_mutation
CD8T
0.00
0.05
0.10
0.15
0.20
FALSE
TRUE
hyper_mutation
Monocyte.E
0.00
0.05
0.10
FALSE
TRUE
hyper_mutation
Monocyte
0.0
0.1
0.2
FALSE
TRUE
hyper_mutation