Analysis of MTHFD2L mRNA levels in scRNA seq

Yizhou Yu

updated: 2023-01-16 ¶

download data:

```
curl -o local.rds "https://corpora-data-prod.s3.amazonaws.com/f541adea-179c-484e-9f06-eaf8d27bca94/loca
```

```
library(Seurat)
```

```
## Attaching SeuratObject
```

```
library(ggplot2)
```

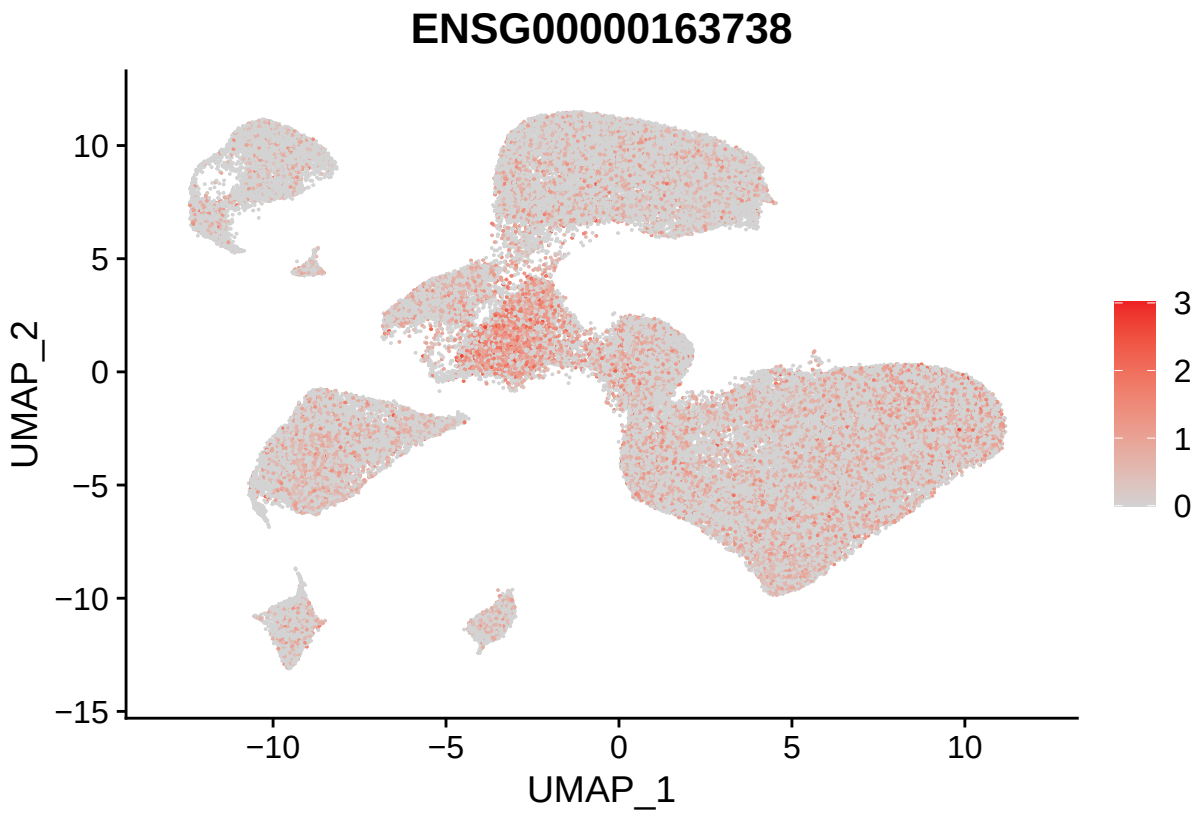
Load seurat dt

```
dt = readRDS("dt/otero.rds")
```

dt is too big.. need to purge

feature plot for MTHFD2L ENSG00000163738

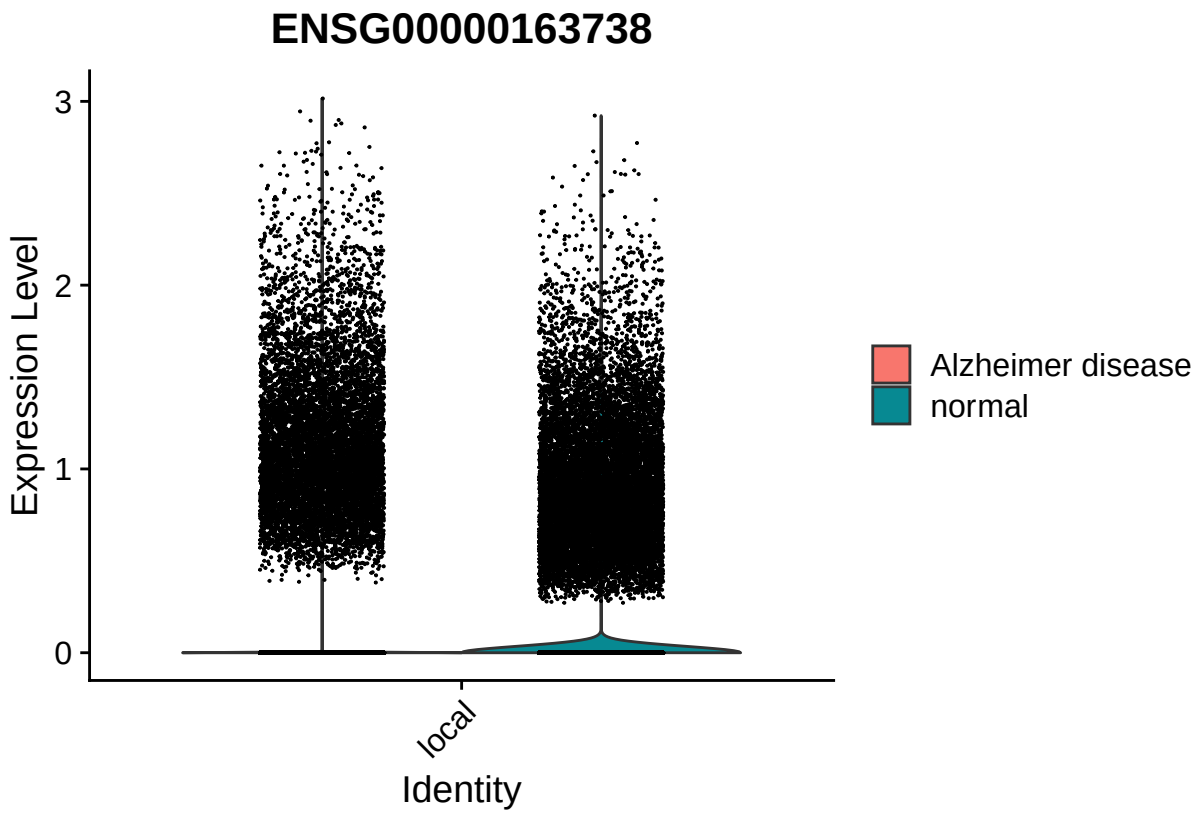
```
FeaturePlot(dt, features = "ENSG00000163738", cols = c("lightgrey","#ED2024"))
```




```
ggsave("fig/MTHFD2L_levels_UMAP.pdf", width = 6, height = 4)
```

```
VlnPlot(object = dt, features = "ENSG00000163738", split.by = "disease")
```

```
## The default behaviour of split.by has changed.
## Separate violin plots are now plotted side-by-side.
## To restore the old behaviour of a single split violin,
## set split.plot = TRUE.
##
## This message will be shown once per session.
```



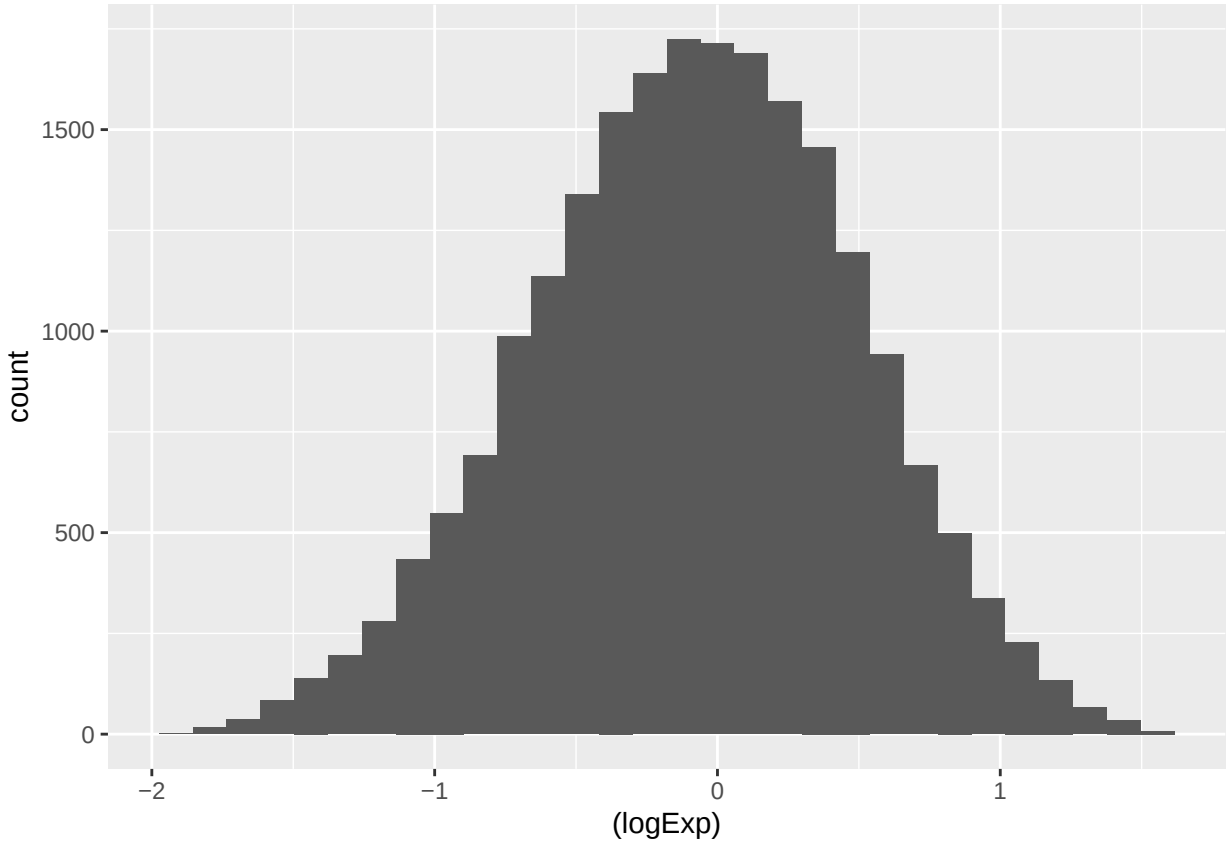
Exploratory analyses and QC

PARP1 levels

```
MTHFD2L_dt = data.frame(cellID = dt@assays$RNA@data@Dimnames[2],
                        mt = dt@meta.data$percent.mt,
                        age = as.numeric(as.character(dt@meta.data$Age)),
                        type = dt@meta.data$Cell.Types,
                        braak = dt@meta.data$Braak,
                        disease = dt@meta.data$disease,
                       sex = dt@meta.data$sex,
                       ethnicity = dt@meta.data$self_reported_ethnicity,
                       patientID = dt@meta.data$donor_id,
                       sort = dt@meta.data$SORT,
                       Exp = GetAssayData(object = dt, slot = "data")["ENSG00000163738",])
MTHFD2L_dt$logExp = log2(MTHFD2L_dt$Exp)
write.csv(MTHFD2L_dt, "dt_out/MTHFD2L_dt_subset.csv", row.names = F)
```

```
MTHFD2L_dt_finite = subset(MTHFD2L_dt, Exp != 0)
ggplot(MTHFD2L_dt_finite, aes(x=(logExp))) + geom_histogram()
```

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



Data needs to be logged

Initial model

```
# ensure that the reference variable is "control patients"
MTHFD2L_dt_finite <- within(MTHFD2L_dt_finite, disease <- relevel(disease, ref = 2))
lm_simplified = lm(data = MTHFD2L_dt_finite, formula = logExp ~ disease + as.numeric(age) + sex)
summary(lm_simplified)
```

```
##
## Call:
## lm(formula = logExp ~ disease + as.numeric(age) + sex, data = MTHFD2L_dt_finite)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65327 -0.37419 -0.00514  0.37638  1.86389
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.1490257  0.0327658   4.548 5.44e-06 ***
## diseaseAlzheimer disease   0.4066557  0.0087456  46.498  < 2e-16 ***
## as.numeric(age)           -0.0056267  0.0004398 -12.794  < 2e-16 ***
## sexmale                    0.0238828  0.0081157   2.943  0.00326 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.5382 on 21336 degrees of freedom
## Multiple R-squared:  0.09721,    Adjusted R-squared:  0.09709
## F-statistic: 765.8 on 3 and 21336 DF,  p-value: < 2.2e-16

                     Estimate Std. Error t value Pr(>|t|)
```

(Intercept) 0.1490257 0.0327658 4.548 5.44e-06 ***diseaseAlzheimer disease 0.4066557 0.0087456 46.498 < 2e-16*** as.numeric(age) -0.0056267 0.0004398 -12.794 < 2e-16 * **sexmale 0.0238828 0.0081157 2.943 0.00326**

Note: lmerTest uses the stringent Satterthwaite approximation, which is based on SAS proc mixed theory.

```
library(lmerTest)
```

```
## Loading required package: lme4

## Loading required package: Matrix

##
## Attaching package: 'lmerTest'

## The following object is masked from 'package:lme4':
##
##     lmer

## The following object is masked from 'package:stats':
##
##     step
```

```
lm_simplified_lmer = lmer(data = MTHFD2L_dt_finite, formula = logExp ~ disease + as.numeric(age) + sex+
summary(lm_simplified_lmer)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: logExp ~ disease + as.numeric(age) + sex + (1 | patientID)
##    Data: MTHFD2L_dt_finite
##
##      AIC      BIC   logLik deviance df.resid
##  30713.6  30761.4 -15350.8  30701.6    21334
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2460 -0.6971 -0.0276  0.6631  3.5776
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  patientID (Intercept) 0.05032  0.2243
##  Residual              0.24578  0.4958
## Number of obs: 21340, groups:  patientID, 16
##
## Fixed effects:
##                           Estimate Std. Error        df t value Pr(>|t|)
```

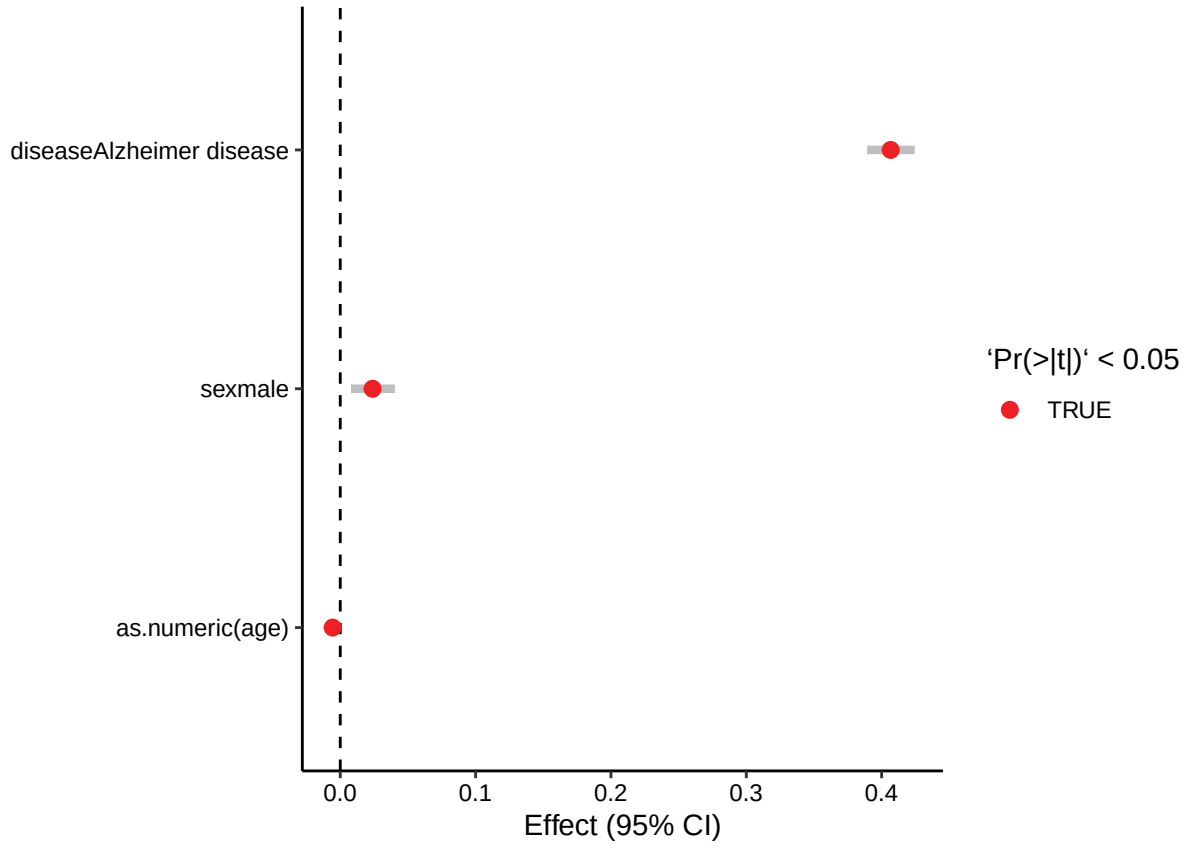
```
## (Intercept)               0.203991   0.463987 16.016326   0.440  0.66607
## diseaseAlzheimer disease  0.367286   0.120857 16.020275   3.039  0.00781 **
## as.numeric(age)          -0.005529   0.006124 16.021111  -0.903  0.37993
## sexmale                   0.063264   0.122019 16.003474   0.518  0.61122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) dssAld as.n()
## dssAlzhmrds  0.120
## as.numrc(g) -0.973 -0.278
## sexmale     -0.448  0.143  0.308
```

```
                    Estimate Std. Error        df t value Pr(>|t|)
```

(Intercept) 0.203991 0.463987 16.016388 0.440 0.66607
diseaseAlzheimer disease 0.367286 0.120857 16.020265 3.039 0.00781 ** as.numeric(age) -0.005529 0.006124 16.021188 -0.903 0.37993
sexmale 0.063264 0.122019 16.003467 0.518 0.61122

```
yy_plt_forest_lm = function(lm_model){
  coef_mat = as.data.frame(summary(lm_model)$coefficients)
  coef_mat = na.omit(coef_mat)
  conf = na.omit(confint.default(lm_model, level = 0.95))
  coef_mat$lci = conf[,1]
  coef_mat$uci = conf[,2]
  coef_mat$label = row.names(coef_mat)
  coef_mat = subset(coef_mat, label != "(Intercept)")
  ggplot(data=coef_mat, aes(x=reorder(label,Estimate),y=Estimate, fill = `Pr(>|t|)` < 0.05, color = `Pr
    geom_hline(yintercept=0, lty=2) +
    geom_errorbar(aes(ymin=lci, ymax=uci),
                  width=0,                    # Width of the error bars
                  position=position_dodge(.9), color = "grey", size = 1.5) +
    geom_point(shape=21, size = 2.5)+
    #geom_pointrange(aes(fill = `Pr(>|t|)` < 0.05)) +
    coord_flip() +  # flip coordinates (puts labels on y axis)
    xlab("") + ylab("Effect (95% CI)") +
    scale_fill_manual(values = c("#ED2024", "#939598"))+
    scale_color_manual(values = c("#ED2024", "#939598"))+
    theme_classic()+
    theme(axis.text.x=element_text(colour="black"),
          axis.text.y=element_text(colour="black"))
}
yy_plt_forest_lm(lm_simplified)
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
```




```
ggsave("fig/MTHFD2L_single_cell_expression.pdf", width = 5, height = 2)
```

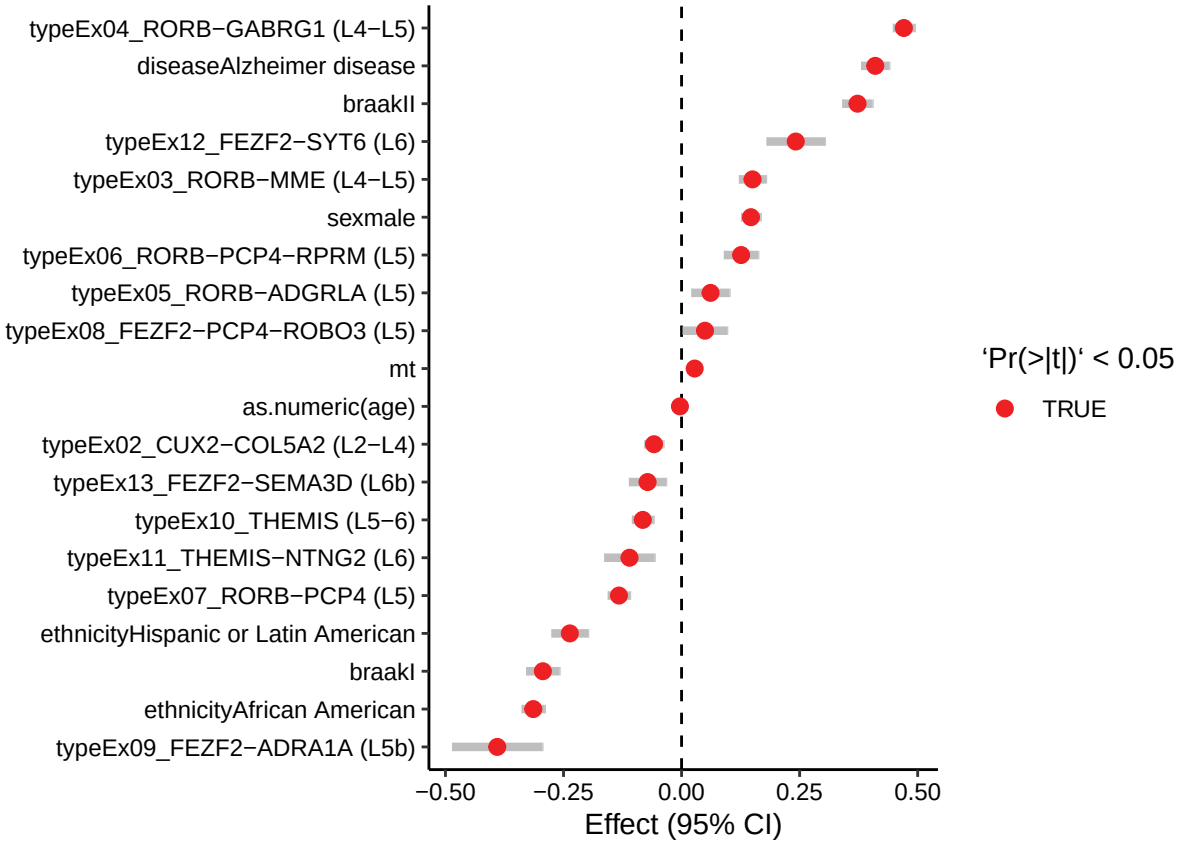
Add more covariates

```
lm = lm(data = MTHFD2L_dt_finite, formula = logExp ~ disease + as.numeric(age) + mt + type + braak + se
summary(lm)
```

```
##
## Call:
## lm(formula = logExp ~ disease + as.numeric(age) + mt + type +
##     braak + sex + ethnicity, data = MTHFD2L_dt_finite)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81625 -0.32575 -0.01179  0.32396  1.91266
##
## Coefficients: (2 not defined because of singularities)
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -0.1244238  0.0328951  -3.782 0.000156 ***
## diseaseAlzheimer disease            0.4101612  0.0153506  26.719  < 2e-16 ***
## as.numeric(age)                    -0.0034697  0.0004335  -8.003 1.28e-15 ***
## mt                                  0.0277863  0.0022177  12.529  < 2e-16 ***
## typeEx02_CUX2-COL5A2 (L2-L4)       -0.0583826  0.0103350  -5.649 1.63e-08 ***
## typeEx03_RORB-MME (L4-L5)           0.1502559  0.0147139  10.212  < 2e-16 ***
## typeEx04_RORB-GABRG1 (L4-L5)        0.4707232  0.0118640  39.677  < 2e-16 ***
```

```
## typeEx05_RORB-ADGRLA (L5)             0.0613912  0.0205959   2.981 0.002879 **
## typeEx06_RORB-PCP4-RPRM (L5)          0.1260969  0.0186494   6.761 1.40e-11 ***
## typeEx07_RORB-PCP4 (L5)              -0.1325890  0.0120932 -10.964  < 2e-16 ***
## typeEx08_FEZF2-PCP4-ROBO3 (L5)        0.0495991  0.0240864   2.059 0.039486 *
## typeEx09_FEZF2-ADRA1A (L5b)          -0.3902621  0.0488426  -7.990 1.42e-15 ***
## typeEx10_THEMIS (L5-6)               -0.0821783  0.0117923  -6.969 3.29e-12 ***
## typeEx11_THEMIS-NTNG2 (L6)           -0.1102046  0.0272055  -4.051 5.12e-05 ***
## typeEx12_FEZF2-SYT6 (L6)              0.2418334  0.0315131   7.674 1.74e-14 ***
## typeEx13_FEZF2-SEMA3D (L6b)          -0.0719653  0.0201603  -3.570 0.000358 ***
## braakI                               -0.2937293  0.0180765 -16.249  < 2e-16 ***
## braakII                               0.3726119  0.0163560  22.781  < 2e-16 ***
## braakVI                                      NA         NA      NA       NA
## sexmale                               0.1469290  0.0106637  13.778  < 2e-16 ***
## ethnicityHispanic or Latin American -0.2366271  0.0197967 -11.953  < 2e-16 ***
## ethnicityAfrican American           -0.3140051  0.0125392 -25.042  < 2e-16 ***
## ethnicityunknown                             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4697 on 21319 degrees of freedom
## Multiple R-squared:  0.3131, Adjusted R-squared:  0.3125
## F-statistic: 485.9 on 20 and 21319 DF,  p-value: < 2.2e-16
```

```
yy_plt_forest_lm(lm)
```



```
ggsave("fig/MTHFD2L_single_cell_expression_full_model.pdf", width = 5, height = 3)
```

```
lm_full_lmer = lmer(data = MTHFD2L_dt_finite, formula = logExp ~ disease + as.numeric(age) + mt + type +
```

```
## fixed-effect model matrix is rank deficient so dropping 2 columns / coefficients
```

```
summary(lm_full_lmer)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: logExp ~ disease + as.numeric(age) + mt + type + braak + sex +
##     ethnicity + (1 | patientID)
##    Data: MTHFD2L_dt_finite
##
##      AIC      BIC   logLik deviance df.resid
##  27276.4  27459.7 -13615.2  27230.4    21317
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.0464 -0.6894 -0.0280  0.6759  4.1451
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  patientID (Intercept) 0.01596  0.1263
##  Residual              0.20904  0.4572
## Number of obs: 21340, groups:  patientID, 16
##
## Fixed effects:
##                                      Estimate Std. Error         df t value
## (Intercept)                         -2.098e-01  2.730e-01  1.599e+01  -0.769
## diseaseAlzheimer disease             4.244e-01  1.489e-01  1.584e+01   2.851
## as.numeric(age)                     -2.542e-03  3.604e-03  1.599e+01  -0.705
## mt                                   3.942e-02  2.343e-03  2.120e+04  16.828
## typeEx02_CUX2-COL5A2 (L2-L4)        -5.628e-02  1.010e-02  2.133e+04  -5.571
## typeEx03_RORB-MME (L4-L5)            1.569e-01  1.436e-02  2.133e+04  10.928
## typeEx04_RORB-GABRG1 (L4-L5)         4.601e-01  1.162e-02  2.133e+04  39.578
## typeEx05_RORB-ADGRLA (L5)            6.324e-02  2.010e-02  2.133e+04   3.147
## typeEx06_RORB-PCP4-RPRM (L5)         1.309e-01  1.821e-02  2.133e+04   7.190
## typeEx07_RORB-PCP4 (L5)             -1.262e-01  1.186e-02  2.133e+04 -10.639
## typeEx08_FEZF2-PCP4-ROBO3 (L5)       7.688e-02  2.351e-02  2.133e+04   3.270
## typeEx09_FEZF2-ADRA1A (L5b)         -3.883e-01  4.757e-02  2.132e+04  -8.163
## typeEx10_THEMIS (L5-6)              -8.922e-02  1.154e-02  2.133e+04  -7.734
## typeEx11_THEMIS-NTNG2 (L6)          -1.324e-01  2.651e-02  2.133e+04  -4.996
## typeEx12_FEZF2-SYT6 (L6)             2.249e-01  3.072e-02  2.133e+04   7.322
## typeEx13_FEZF2-SEMA3D (L6b)         -7.805e-02  1.966e-02  2.133e+04  -3.971
## braakI                              -2.780e-01  1.807e-01  1.582e+01  -1.538
## braakII                              3.595e-01  1.572e-01  1.584e+01   2.288
## sexmale                              1.348e-01  8.923e-02  1.596e+01   1.511
## ethnicityHispanic or Latin American -2.182e-01  2.033e-01  1.580e+01  -1.073
## ethnicityAfrican American           -2.357e-01  1.103e-01  1.595e+01  -2.138
##                                     Pr(>|t|)
## (Intercept)                          0.45338
```

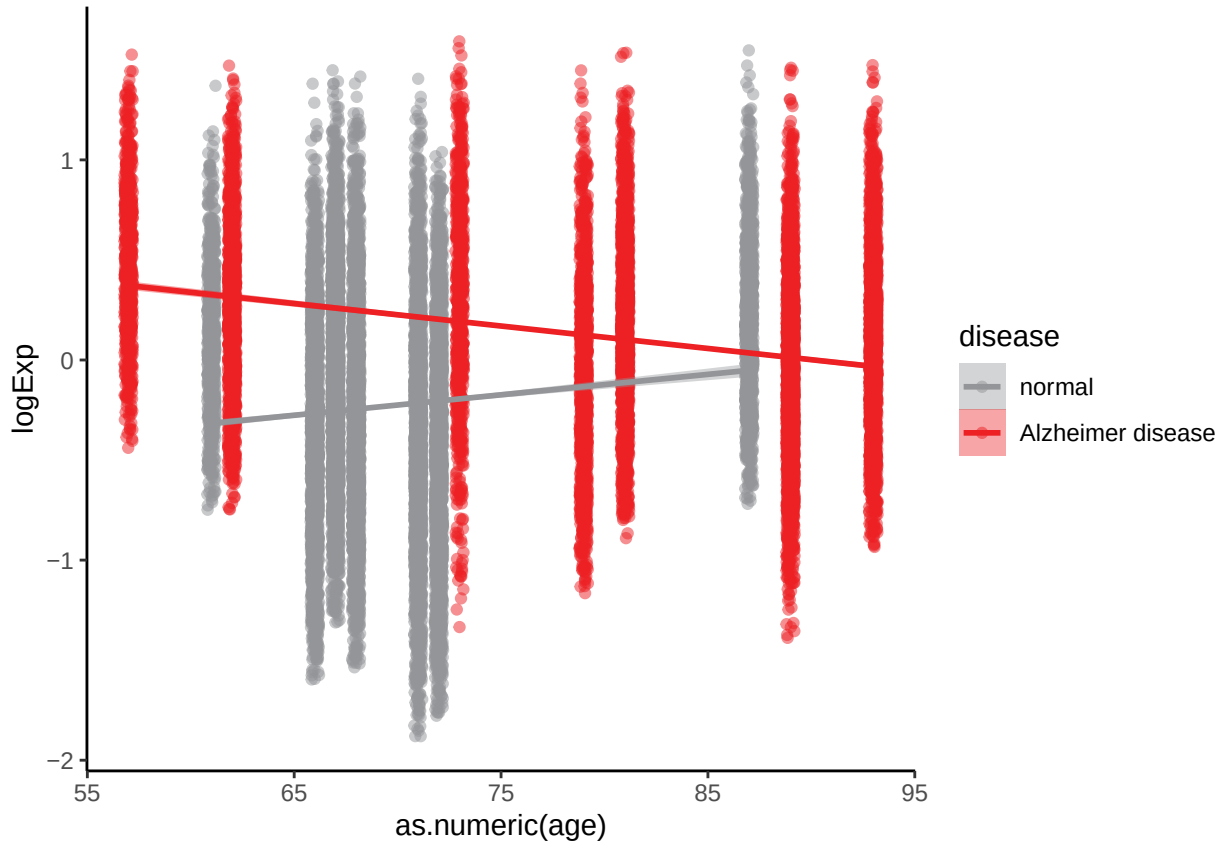
```
## diseaseAlzheimer disease             0.01165 *
## as.numeric(age)                      0.49085
## mt                                   < 2e-16 ***
## typeEx02_CUX2-COL5A2 (L2-L4)        2.56e-08 ***
## typeEx03_RORB-MME (L4-L5)            < 2e-16 ***
## typeEx04_RORB-GABRG1 (L4-L5)         < 2e-16 ***
## typeEx05_RORB-ADGRLA (L5)            0.00165 **
## typeEx06_RORB-PCP4-RPRM (L5)        6.71e-13 ***
## typeEx07_RORB-PCP4 (L5)              < 2e-16 ***
## typeEx08_FEZF2-PCP4-ROBO3 (L5)       0.00108 **
## typeEx09_FEZF2-ADRA1A (L5b)         3.46e-16 ***
## typeEx10_THEMIS (L5-6)              1.08e-14 ***
## typeEx11_THEMIS-NTNG2 (L6)          5.91e-07 ***
## typeEx12_FEZF2-SYT6 (L6)            2.53e-13 ***
## typeEx13_FEZF2-SEMA3D (L6b)         7.19e-05 ***
## braakI                               0.14369
## braakII                              0.03625 *
## sexmale                              0.15038
## ethnicityHispanic or Latin American  0.29938
## ethnicityAfrican American            0.04838 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 21 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it

## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 2 columns / coefficients
```

```
ggplot(MTHFD2L_dt_finite, aes(x=as.numeric(age), y=logExp, color=disease, fill = disease)) +
  geom_point(position = position_jitter(seed = 1, width = 0.2),
             alpha = 0.5)+
  geom_smooth(method = "lm")+
  scale_fill_manual(values=c("#939598","#ED2024"))+
  scale_color_manual(values=c("#939598","#ED2024"))+
  theme_classic()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

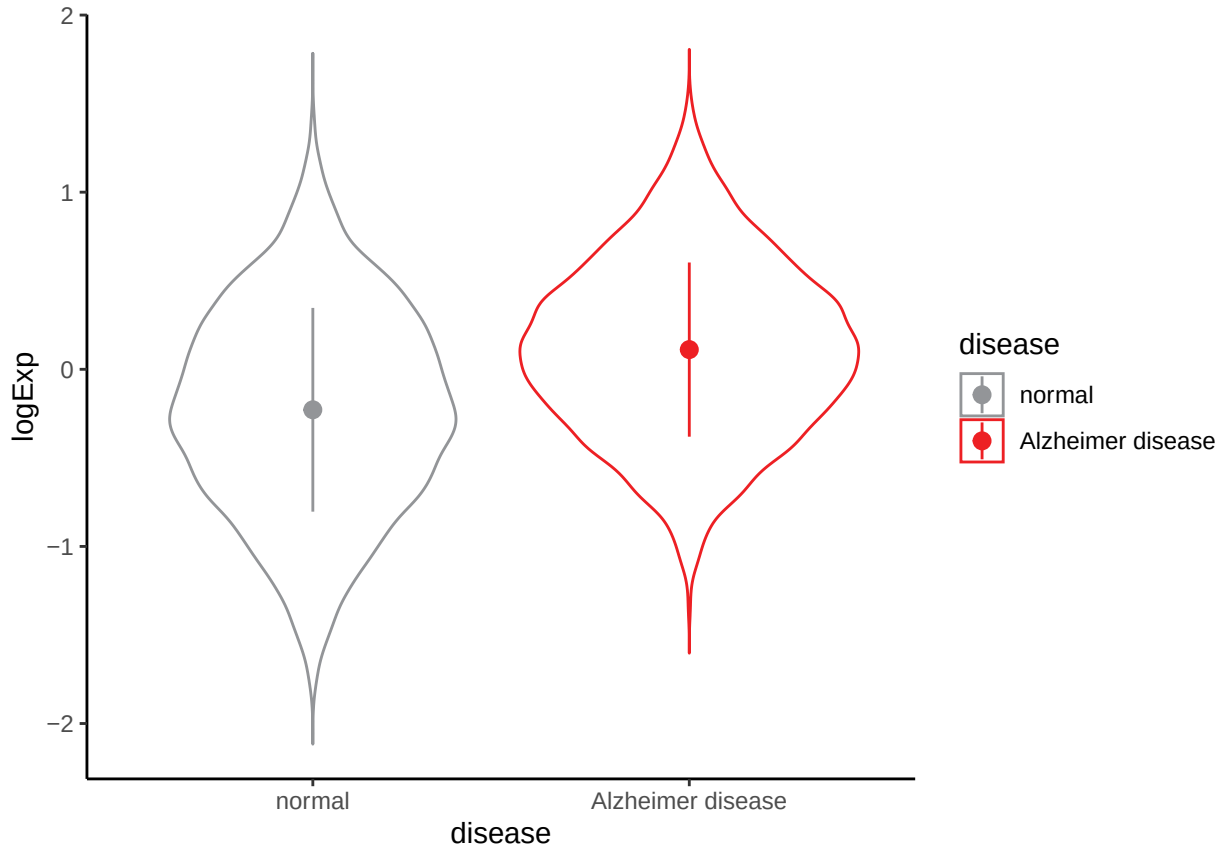



```
ggsave("fig/MTHFD2L_single_cell_scatter_age_disease.pdf", width = 5, height = 3)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
data_summary <- function(x) {
   m <- mean(x)
   ymin <- m-sd(x)
   ymax <- m+sd(x)
   return(c(y=m,ymin=ymin,ymax=ymax))
}

ggplot(MTHFD2L_dt_finite, aes(x=disease, y=logExp, color=disease)) +
    geom_violin(trim = F) +
  #geom_point(position = position_jitter(seed = 1, width = 0.2), alpha = 0.5)+
  stat_summary(fun.data=data_summary)+
  scale_fill_manual(values=c("#939598","#ED2024"))+
  scale_color_manual(values=c("#939598","#ED2024"))+
  theme_classic()
```




```
ggsave("fig/MTHFD2L_expression_violin.pdf", width = 4, height = 3)
```

Add cell numbers

```
summary(MTHFD2L_dt_finite)
```

```
##  c..C0001_AAACCTGAGGAGTTGC.1....C0001_AAACCTGTCAATCTCT.1....C0001_AAACCTGTCGTCACGG.1...
##  Length:21340
##  Class :character
##  Mode  :character
##
##
##
##
##        mt               age                                type      braak
##  Min.   : 0.0000   Min.   :57.0   Ex02_CUX2-COL5A2 (L2-L4):4860   0 :3505
##  1st Qu.: 0.7465   1st Qu.:67.0   Ex01_CUX2-LAMP5 (L2-L3) :3754   I :3775
##  Median : 1.4706   Median :71.0   Ex10_THEMIS (L5-6)      :2849   II:4887
##  Mean   : 1.8233   Mean   :74.2   Ex04_RORB-GABRG1 (L4-L5):2782   VI:9173
##  3rd Qu.: 2.4483   3rd Qu.:81.0   Ex07_RORB-PCP4 (L5)     :2589
##  Max.   :20.4033   Max.   :93.0   Ex03_RORB-MME (L4-L5)   :1405
##                                   (Other)                 :3101
##               disease          sex                                ethnicity
##  normal           :12167   female:11333   European                      : 3775
##  Alzheimer disease: 9173   male  :10007   Hispanic or Latin American: 1966
```

```
##                                         African American          : 2945
##                                         unknown                   :12654
##
##
##
##     patientID             sort            Exp             logExp
##  Subject7:2440   AT8        : 3530   Min.   :0.2717   Min.   :-1.87986
##  CTRL-7  :2264   MAP2       : 5643   1st Qu.:0.7245   1st Qu.:-0.46497
##  CTRL-8  :1966   MAP2control:12167   Median :0.9534   Median :-0.06892
##  CTRL-6  :1959                      Mean   :1.0187   Mean   :-0.08246
##  CTRL-5  :1816                      3rd Qu.:1.2461   3rd Qu.: 0.31745
##  Subject1:1616                      Max.   :3.0155   Max.   : 1.59238
##  (Other) :9279
```

normal :12167 Alzheimer disease: 9173