This script includes the data analysis of the gene age data together with the analysed co-factors:

co.factors[[1]] <- protein intrinsic disorder  
co.factors[[2]] <- gene expression  
co.factors[[3]] <- protein length  
co.factors[[4]] <- mean relative solvent accessability per gene  

The first chunk of the script will show the correlations between gene age and each of the co-factors.
# Gene Age vs. Protein Length

length.age <- lapply(data.list, function(x) {
  na.omit(unique(subset(x, select = c("GeneID", "Clade", "Length", "species"))))
})

# saving the table S2 Data
length_age_df <- rbindlist(length.age)
write.table(length_age_df, file = ".../Co-factors_continuous/S2_Data.csv", sep = "\t",
  col.names = T, row.names = F, quote = F)

# summarizing protein length data for each clade
fun.q <- function(x) {
  q <- quantile(x$Length, na.rm = TRUE)
  firstQ <- q[[2]]
  thirdQ <- q[[4]]
  med.length <- q[[3]]
  tbl <- data.frame(med.length, firstQ, thirdQ)
}

sum.length.age <- lapply(length.age, function(x) {
  ddply(x, c("Clade", "species"), fun.q)
})

# plotting each of the output tables
plot.age.length <- lapply(sum.length.age, function(x) {
  ggplot(x, aes(x = Clade, y = med.length)) +
  geom_line(col = "black", size = 0.2)+
  geom_ribbon(aes(ymax=thirdQ, ymin=firstQ), alpha=0.2) +
  geom_point(size=.2)+
  ylab("") +
  xlab("Clade") +
  #scale_x_sqrt() +
  scale_fill_grey() +
  scale_color_grey() +
  ggtitle(as.character(x$species)) +
  theme_bw() +
  theme.plot()
})
plot.age.length
Table 1: Correlation between gene age and protein length in Drosophila

<table>
<thead>
<tr>
<th>species</th>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>tau</td>
<td>-0.4848485</td>
<td>0.0282123</td>
</tr>
</tbody>
</table>

# estimating the correlations
stats.age.length <- lapply(sum.length.age, function(x) {
  cor.df <- cor.test(x$Clade, x$med.length, method = "kendall", exact = FALSE)
  species <- unique(x$species)
  Kendall.tau <- cor.df$estimate
  p.value <- cor.df$p.value
  data.frame(species, Kendall.tau, p.value)
})

# showing the tables
for(i in stats.age.length) {
  print(kable(x = i, caption = paste0("Correlation between gene age and protein length in ", unique(i$species))))
}

# Gene Age vs. Gene Expression
exp.age <- lapply(data.list, function(x) {
Table 2: Correlation between gene age and protein length in Arabidopsis

<table>
<thead>
<tr>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.847619</td>
<td>1.06e-05</td>
</tr>
</tbody>
</table>

```r
na.omit(unique(subset(x, select = c("GeneID", "Clade", "ExpressionMean", "species"))))
```

```r
# saving the table S3 Data
exp_age_df <- rbindlist(exp.age)
write.table(exp_age_df, file = "../Co-factors_continuous/S3_Data.csv", sep = "\t",
            col.names = T, row.names = F, quote = F)

# summarizing protein length data for each clade
fun.q <- function(x) {
  q <- quantile(x$ExpressionMean, na.rm = TRUE)
  firstQ <- q[[2]]
  thirdQ <- q[[4]]
  med.exp <- q[[3]]
  tbl <- data.frame(med.exp, firstQ, thirdQ)
}

sum.exp.age <- lapply(exp.age,
                      function(x) {
                        ddply(x, c("Clade", "species"), fun.q)
                      })

# plotting each of the output tables
plot.age.exp <- lapply(sum.exp.age,
                       function(x) {
                         ggplot(x, aes(x = Clade, y = med.exp)) +
                         geom_line(col = "black", size = 0.2) +
                         geom_ribbon(aes(ymax=thirdQ, ymin=firstQ), alpha=0.2) +
                         geom_point(size=.2) +
                         ylab("") +
                         xlab("Clade") +
                         #scale_x_sqrt() +
                         scale_fill_grey() +
                         scale_color_grey() +
                         ggtitle(as.character(x$species)) +
                         theme_bw() +
                         theme.plot()
                       })

plot.age.exp
```
Table 3: Correlation between gene age and gene expression in Drosophila

<table>
<thead>
<tr>
<th>species</th>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>tau Drosophila</td>
<td>-0.5954372</td>
<td>0.0073482</td>
</tr>
</tbody>
</table>

```
# estimating the correlations
stats.age.exp <- lapply(sum.exp.age, function(x) {
  cor.df <- cor.test(x$Clade, x$med.exp, method = "kendall", exact = FALSE)
  species <- unique(x$species)
  Kendall.tau <- cor.df$estimate
  p.value <- cor.df$p.value
  data.frame(species, Kendall.tau, p.value)
})

# showing the tables
for(i in stats.age.exp) {
  print(kable(x = i, caption = paste0("Correlation between gene age 
and gene expression in ", unique(i$species))))
}
```

```[2]
```
```
Table 4: Correlation between gene age and gene expression in Arabidopsis

<table>
<thead>
<tr>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.7904762</td>
<td>4e-05</td>
</tr>
</tbody>
</table>

```r
na.omit(unique(subset(x, select = c("GeneID", "Clade", "mean.rsa", "species"))))

# saving the table S3 Data
rsa_age_df <- rbindlist(rsa.age)
write.table(rsa_age_df, file = "./Co-factors_continuous/S4_Data.csv", sep = "\t",
            col.names = T, row.names = F, quote = F)

# summarizing protein length data for each clade
fun.q <- function(x) {  
  q <- quantile(x$mean.rsa, na.rm = TRUE)
  firstQ <- q[[2]]
  thirdQ <- q[[4]]
  med.rsa <- q[[3]]
  tbl <- data.frame(med.rsa, firstQ, thirdQ)
}

sum.rsa.age <- lapply(rsa.age, function(x) {  
  ddply(x, c("Clade", "species"), fun.q)
})

# plotting each of the output tables
plot.age.rsa <- lapply(sum.rsa.age, function(x) {  
  ggplot(x, aes(x = Clade, y = med.rsa)) +
  geom_line(col = "black", size = 0.2) +
  geom_ribbon(aes(ymax=thirdQ, ymin=firstQ), alpha=0.2) +
  geom_point(size=.2) +
  ylab("") +
  xlab("Clade") +
  #scale_x_sqrt() +
  scale_fill_grey() +
  scale_color_grey() +
  ggtitle(as.character(x$species)) +
  theme_bw() +
  theme.plot()
})
plot.age.rsa
```
Drosophila

Clade

[[1]]
Table 5: Correlation between gene age and rsa in Drosophila

<table>
<thead>
<tr>
<th>species</th>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>tau</td>
<td>0.6363636</td>
<td>0.0039762</td>
</tr>
</tbody>
</table>

```
# estimating the correlations
stats.age.rsa <- lapply(sum.rsa.age, function(x) {
  cor.df <- cor.test(x$med.rsa, x$Clade, method = "kendall", exact = FALSE)
  species <- unique(x$species)
  Kendall.tau <- cor.df$estimate
  p.value <- cor.df$p.value
  data.frame(species, Kendall.tau, p.value)
})

# showing the tables
for(i in stats.age.rsa) {
  print(kable(x = i, caption = paste0("Correlation between gene age and rsa in ", unique(i$species))))
}
```

```
# Gene Age vs. Protein Intrinsic Disorder
disorder.age <- lapply(data.list, function(x) {
```

### Arabidopsis

![Graph showing correlation between gene age and RSA in Arabidopsis](image_url)
Table 6: Correlation between gene age and rsa in Arabidopsis

<table>
<thead>
<tr>
<th>species</th>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>0.6952381</td>
<td>0.0003032</td>
</tr>
</tbody>
</table>

```r
na.omit(unique(subset(x, select = c("GeneID", "Clade", "mean.dis", "species"))))
```

```r
# saving the table S3 Data
dis_age_df <- rbindlist(disorder.age)
write.table(dis_age_df, file = "../Co-factors_continuous/S5_Data.csv", sep = "\t",
col.names = T, row.names = F, quote = F)
```

```r
# summarizing protein length data for each clade
fun.q <- function(x) {
  q <- quantile(x$mean.dis, na.rm = TRUE)
  firstQ <- q[[2]]
  thirdQ <- q[[4]]
  med.dis <- q[[3]]
  tbl <- data.frame(med.dis, firstQ, thirdQ)
}
sum.dis.age <- lapply(disorder.age, function(x) {
  ddply(x, c("Clade", "species"), fun.q)
})

# plotting each of the output tables
plot.age.dis <- lapply(sum.dis.age, function(x) {
  ggplot(x, aes(x = Clade, y = med.dis)) +
  geom_line(col = "black", size = 0.2)+
  geom_ribbon(aes(ymax=thirdQ, ymin=firstQ), alpha=0.2) +
  geom_point(size=.2)+
ylab("") +
xlab("Clade") +
#scale_x_sqrt() +
scale_fill_grey() +
scale_color_grey() +
ggtitle(as.character(x$species)) +
theme_bv() +
theme.plot()
})
plot.age.dis
```
Table 7: Correlation between gene age and disorder in Drosophila

<table>
<thead>
<tr>
<th>species</th>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drosophila</td>
<td>0.6060606</td>
<td>0.0060899</td>
</tr>
</tbody>
</table>

Arabidopsis

```
# estimating the correlations
stats.age.dis <- lapply(sum.dis.age, function(x) {
  cor.df <- cor.test(x$med.dis, x$Clade, method = "kendall", exact = FALSE)
  species <- unique(x$species)
  Kendall.tau <- cor.df$estimate
  p.value <- cor.df$p.value
  data.frame(species, Kendall.tau, p.value)
})

# showing the tables
for(i in stats.age.dis) {
  print(kable(x = i, caption = paste0("Correlation between gene age and disorder in ", unique(i$species))))
}
```

setwd("~/Dropbox/SupplementaryData_GeneAge/Co-factors_continuous/grapes_analysis/")

# change here to the respective folder where you keep the data

# calling all output tables
Table 8: Correlation between gene age and disorder in Arabidopsis

<table>
<thead>
<tr>
<th>tau</th>
<th>species</th>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>0.4666667</td>
<td>0.0153138</td>
<td></td>
</tr>
</tbody>
</table>

co.factors <- list.files(".", ".csv")

# reading each table into a list
tbl.list <- lapply(co.factors, read.table, header = TRUE)

# arranging the table for plotting:
tbl.list2 <- lapply(tbl.list, function(x) {
melt(x, id.vars = c("GeneAge", "category", "co_factor", "species"),
  measure.vars = c("dnds", "omegaNA", "omegaA"))
})

# function to estimate the mean and standard deviation to plot the results with the
# mean of the bootstrap replicates and the 95% confidence interval
fun <- function(x){
c(mean=mean(x), sd=sd(x))
}

# applying the above function to each output table for each value of each estimate
# (dnds, omegaA, omegaNA) for each value of the variable being analyzed for each species
tbl.sum <- lapply(tbl.list2, function(x) {
  summaryBy(value ~ variable + GeneAge + category + species + co_factor, data=x, FUN = fun)
})

# to change the estimate name to the respective symbol
tbl.sum2 <- lapply(tbl.sum, function(x) {
  ddply(x, c("GeneAge", "category", "co_factor", "species"), function(y) {
    y$variable <- factor(y$variable, levels = c("dnds", "omegaNA", "omegaA"))
    levels(y$variable) <- c(expression(omega), expression(omega[na]),
                           expression(omega[a]))
    return(y)
  })
})

Including Plots

In the next chunk the script to plot the results is represented. The same order will follow.

# plotting each of the output tables
plot.co_factors <- lapply(tbl.sum2, function(x) {
ggplot(x, aes(x = GeneAge, y = value.mean, fill = category)) +
geom_line(col = "black", size = 0.2)+
geom_ribbon(aes(ymin=value.mean + 1.96*value.sd,
  ymax=value.mean - 1.96*value.sd, fill = category), alpha=0.6) +
geom_point(size=.2)+
facet_grid(species~variable, scales = "free_x", labeller = label_parsed) +
ylab("") +
```r
xlab("Gene Age") +
#scale_x_sqrt() +
scale_fill_grey() +
scale_color_grey() +
theme_bw() +
ggtitle(as.character(x$co_factor)) +
theme.plot()
}

plot.co_factors
```

![Graph showing gene age and length](image)
Gene Age
category
High
Low
Expression

Arabidopsis
Drosophila

Gene Age
The last section shows how the statistical analyses were performed.

```r
tbl.stat <- lapply(tbl.sum2, function(x) {
  ddply(x, c("co_factor", "species", "variable", "category"), function(x) {
    var <- as.numeric(factor(x$GeneAge))
    variable.value <- as.numeric(factor(x$value.mean))
    corr = cor.test(var, variable.value, method = "kendall", exact = FALSE)
    Kendall.tau = corr$estimate
    p.value = corr$p.value
    dat = data.frame(Kendall.tau, p.value)
  })
})

# showing the tables
for(i in tbl.stat) {
  print(kable(x = i, caption = paste0("Statistics for ", unique(i$co_factor))))
}
```
### Table 9: Statistics for Length

<table>
<thead>
<tr>
<th>co_factor</th>
<th>species</th>
<th>variable</th>
<th>category</th>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>Arabidopsis</td>
<td>omega</td>
<td>long</td>
<td>1.0000000</td>
<td>0.0048322</td>
</tr>
<tr>
<td>Length</td>
<td>Arabidopsis</td>
<td>omega</td>
<td>short</td>
<td>1.0000000</td>
<td>0.0048322</td>
</tr>
<tr>
<td>Length</td>
<td>Arabidopsis</td>
<td>omega[na]</td>
<td>long</td>
<td>0.8666667</td>
<td>0.0145950</td>
</tr>
<tr>
<td>Length</td>
<td>Arabidopsis</td>
<td>omega[na]</td>
<td>short</td>
<td>0.8666667</td>
<td>0.0145950</td>
</tr>
<tr>
<td>Length</td>
<td>Arabidopsis</td>
<td>omega[a]</td>
<td>long</td>
<td>-0.2000000</td>
<td>0.5730251</td>
</tr>
<tr>
<td>Length</td>
<td>Arabidopsis</td>
<td>omega[a]</td>
<td>short</td>
<td>0.6000000</td>
<td>0.0908739</td>
</tr>
<tr>
<td>Length</td>
<td>Drosophila</td>
<td>omega</td>
<td>long</td>
<td>0.8666667</td>
<td>0.0145950</td>
</tr>
<tr>
<td>Length</td>
<td>Drosophila</td>
<td>omega</td>
<td>short</td>
<td>0.7333333</td>
<td>0.0387775</td>
</tr>
<tr>
<td>Length</td>
<td>Drosophila</td>
<td>omega[na]</td>
<td>long</td>
<td>0.6000000</td>
<td>0.0908739</td>
</tr>
<tr>
<td>Length</td>
<td>Drosophila</td>
<td>omega[na]</td>
<td>short</td>
<td>0.8666667</td>
<td>0.0145950</td>
</tr>
</tbody>
</table>

### Table 10: Statistics for Expression

<table>
<thead>
<tr>
<th>co_factor</th>
<th>species</th>
<th>variable</th>
<th>category</th>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>Arabidopsis</td>
<td>omega</td>
<td>High</td>
<td>0.8666667</td>
<td>0.0145950</td>
</tr>
<tr>
<td>Expression</td>
<td>Arabidopsis</td>
<td>omega</td>
<td>Low</td>
<td>0.8666667</td>
<td>0.0145950</td>
</tr>
<tr>
<td>Expression</td>
<td>Arabidopsis</td>
<td>omega[na]</td>
<td>High</td>
<td>0.8666667</td>
<td>0.0145950</td>
</tr>
<tr>
<td>Expression</td>
<td>Arabidopsis</td>
<td>omega[na]</td>
<td>Low</td>
<td>1.0000000</td>
<td>0.0048322</td>
</tr>
<tr>
<td>Expression</td>
<td>Arabidopsis</td>
<td>omega[a]</td>
<td>High</td>
<td>0.4666667</td>
<td>0.188460</td>
</tr>
<tr>
<td>Expression</td>
<td>Arabidopsis</td>
<td>omega[a]</td>
<td>Low</td>
<td>0.3333333</td>
<td>0.347558</td>
</tr>
<tr>
<td>Expression</td>
<td>Drosophila</td>
<td>omega</td>
<td>High</td>
<td>0.8666667</td>
<td>0.0145950</td>
</tr>
<tr>
<td>Expression</td>
<td>Drosophila</td>
<td>omega</td>
<td>Low</td>
<td>0.8666667</td>
<td>0.0145950</td>
</tr>
<tr>
<td>Expression</td>
<td>Drosophila</td>
<td>omega[na]</td>
<td>High</td>
<td>1.0000000</td>
<td>0.0048322</td>
</tr>
<tr>
<td>Expression</td>
<td>Drosophila</td>
<td>omega[na]</td>
<td>Low</td>
<td>0.7333333</td>
<td>0.0387775</td>
</tr>
<tr>
<td>Expression</td>
<td>Drosophila</td>
<td>omega[a]</td>
<td>High</td>
<td>0.6000000</td>
<td>0.0908739</td>
</tr>
<tr>
<td>Expression</td>
<td>Drosophila</td>
<td>omega[a]</td>
<td>Low</td>
<td>1.0000000</td>
<td>0.0048322</td>
</tr>
</tbody>
</table>

### Table 11: Statistics for RSA

<table>
<thead>
<tr>
<th>co_factor</th>
<th>species</th>
<th>variable</th>
<th>category</th>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSA</td>
<td>Arabidopsis</td>
<td>omega</td>
<td>High</td>
<td>0.9444444</td>
<td>0.0003930</td>
</tr>
<tr>
<td>RSA</td>
<td>Arabidopsis</td>
<td>omega</td>
<td>Low</td>
<td>1.0000000</td>
<td>0.0001746</td>
</tr>
<tr>
<td>RSA</td>
<td>Arabidopsis</td>
<td>omega[na]</td>
<td>High</td>
<td>0.8888889</td>
<td>0.0008492</td>
</tr>
<tr>
<td>RSA</td>
<td>Arabidopsis</td>
<td>omega[na]</td>
<td>Low</td>
<td>0.7777778</td>
<td>0.0035093</td>
</tr>
<tr>
<td>RSA</td>
<td>Arabidopsis</td>
<td>omega[a]</td>
<td>High</td>
<td>0.7222222</td>
<td>0.0067144</td>
</tr>
<tr>
<td>RSA</td>
<td>Arabidopsis</td>
<td>omega[a]</td>
<td>Low</td>
<td>0.6666667</td>
<td>0.0123434</td>
</tr>
<tr>
<td>RSA</td>
<td>Drosophila</td>
<td>omega</td>
<td>High</td>
<td>0.6363636</td>
<td>0.0064351</td>
</tr>
<tr>
<td>RSA</td>
<td>Drosophila</td>
<td>omega</td>
<td>Low</td>
<td>0.6363636</td>
<td>0.0064351</td>
</tr>
<tr>
<td>RSA</td>
<td>Drosophila</td>
<td>omega[na]</td>
<td>High</td>
<td>0.6727273</td>
<td>0.0039711</td>
</tr>
<tr>
<td>RSA</td>
<td>Drosophila</td>
<td>omega[na]</td>
<td>Low</td>
<td>0.4909091</td>
<td>0.0355579</td>
</tr>
<tr>
<td>RSA</td>
<td>Drosophila</td>
<td>omega[a]</td>
<td>High</td>
<td>0.5636364</td>
<td>0.0158068</td>
</tr>
<tr>
<td>RSA</td>
<td>Drosophila</td>
<td>omega[a]</td>
<td>Low</td>
<td>0.5636364</td>
<td>0.0158068</td>
</tr>
</tbody>
</table>
Table 12: Statistics for Disorder

<table>
<thead>
<tr>
<th>co_factor</th>
<th>species</th>
<th>variable</th>
<th>category</th>
<th>Kendall.tau</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disorder</td>
<td>Arabidopsis</td>
<td>omega</td>
<td>high</td>
<td>1.0000000</td>
<td>0.0000060</td>
</tr>
<tr>
<td>Disorder</td>
<td>Arabidopsis</td>
<td>omega</td>
<td>low</td>
<td>0.9696970</td>
<td>0.0000114</td>
</tr>
<tr>
<td>Disorder</td>
<td>Arabidopsis</td>
<td>omega[na]</td>
<td>high</td>
<td>0.9393939</td>
<td>0.0000212</td>
</tr>
<tr>
<td>Disorder</td>
<td>Arabidopsis</td>
<td>omega[na]</td>
<td>low</td>
<td>0.9090909</td>
<td>0.0000388</td>
</tr>
<tr>
<td>Disorder</td>
<td>Arabidopsis</td>
<td>omega[a]</td>
<td>high</td>
<td>0.6363636</td>
<td>0.0039762</td>
</tr>
<tr>
<td>Disorder</td>
<td>Arabidopsis</td>
<td>omega[a]</td>
<td>low</td>
<td>0.4545455</td>
<td>0.0396693</td>
</tr>
<tr>
<td>Disorder</td>
<td>Drosophila</td>
<td>omega</td>
<td>high</td>
<td>0.6969697</td>
<td>0.0016086</td>
</tr>
<tr>
<td>Disorder</td>
<td>Drosophila</td>
<td>omega</td>
<td>low</td>
<td>0.6363636</td>
<td>0.0039762</td>
</tr>
<tr>
<td>Disorder</td>
<td>Drosophila</td>
<td>omega[na]</td>
<td>high</td>
<td>0.3030303</td>
<td>0.1702344</td>
</tr>
<tr>
<td>Disorder</td>
<td>Drosophila</td>
<td>omega[na]</td>
<td>low</td>
<td>0.5757576</td>
<td>0.0091672</td>
</tr>
<tr>
<td>Disorder</td>
<td>Drosophila</td>
<td>omega[a]</td>
<td>high</td>
<td>0.5151515</td>
<td>0.0197288</td>
</tr>
<tr>
<td>Disorder</td>
<td>Drosophila</td>
<td>omega[a]</td>
<td>low</td>
<td>0.2727273</td>
<td>0.2170890</td>
</tr>
</tbody>
</table>