EpigeneticAge_Statistical_Analysis

R Markdown

This markdown provides a detailed statistical analysis of delta age across different subtypes. It begins by performing t-tests for all three epigenetic clocks—Horvath, PhenoAge, and Hannum—in relation to the subtypes. Scatter plots are then presented to illustrate cross-clock differences and consensus patterns. In addition, box plots of delta age by subtype are included to highlight key separation trends. Finally, a linear model is fitted to further explore these patterns and identify the most suitable clock for downstream analysis.

```
delta_age_with_chrono_age = read.csv("/mnt/home/fayyaz/MPIB-SRT/1040-pelotas/private/data/autobids/delta
clinical_data = read.csv("/mnt/home/fayyaz/MPIB-SRT/1040-pelotas/private/data/autobids/clinical_data.csv
```

```
# t-tests
#t_results <- lapply(delta_vars, function(var) {
#  formula <- as.formula(paste(var, "~ Subtype_binary"))
# t.test(formula, data = delta_age_with_chrono_age)
# })
```

```
# Run the t-tests
# Run the t-tests
t1 <- t.test(delta_Horvath ~ Subtype_binary, data = delta_age_with_chrono_age)
t2 <- t.test(delta_Hannum ~ Subtype_binary, data = delta_age_with_chrono_age)
t3 <- t.test(delta_PhenoAge ~ Subtype_binary, data = delta_age_with_chrono_age)

# Function to return a summary list (instead of printing with cat)
ttest_summary <- function(t_result) {
  list(
    t_statistic = round(t_result$statistic, 3),
    df = round(t_result$parameter, 2),
    p_value = signif(t_result$p.value, 4),
    conf_int = round(t_result$conf.int, 3),
    group_means = round(t_result$estimate, 3)
  )
}

# Print results for each test using knitr-friendly output
results_horvath <- ttest_summary(t1)
results_hannum <- ttest_summary(t2)
results_pheno <- ttest_summary(t3)

# Display as printed outputs in Rmd
results_horvath
```

```
## $t_statistic
##       t
## -10.851
##
## $df
##     df
```

```
## 208.72
##
## $p_value
## [1] 4.831e-22
##
## $conf_int
## [1] -23.376 -16.189
## attr(,"conf.level")
## [1] 0.95
##
## $group_means
##   mean in group Basal mean in group Luminal
##                -8.376                11.407
```

```
results_hannum
```

```
## $t_statistic
##       t
## -10.976
##
## $df
##     df
## 184.05
##
## $p_value
## [1] 6.934e-22
##
## $conf_int
## [1] -24.970 -17.361
## attr(,"conf.level")
## [1] 0.95
##
## $group_means
##   mean in group Basal mean in group Luminal
##                -3.193                17.973
```

```
results_pheno
```

```
## $t_statistic
##      t
## -6.334
##
## $df
##     df
## 189.48
##
## $p_value
## [1] 1.696e-09
##
## $conf_int
## [1] -23.557 -12.369
## attr(,"conf.level")
## [1] 0.95
##
## $group_means
##   mean in group Basal mean in group Luminal
```

```
##                 12.038                 30.001
```

#Interpretation

All three Δ-age clocks show highly significant differences between the Basal and Luminal subtypes.

Δ-Hannum and Δ-Horvath show younger biological age in Basal, while Δ-PhenoAge also shows significantly younger age in Basal but with generally higher values.

The effect size appears largest for Hannum and Horvath (based on means and confidence intervals).

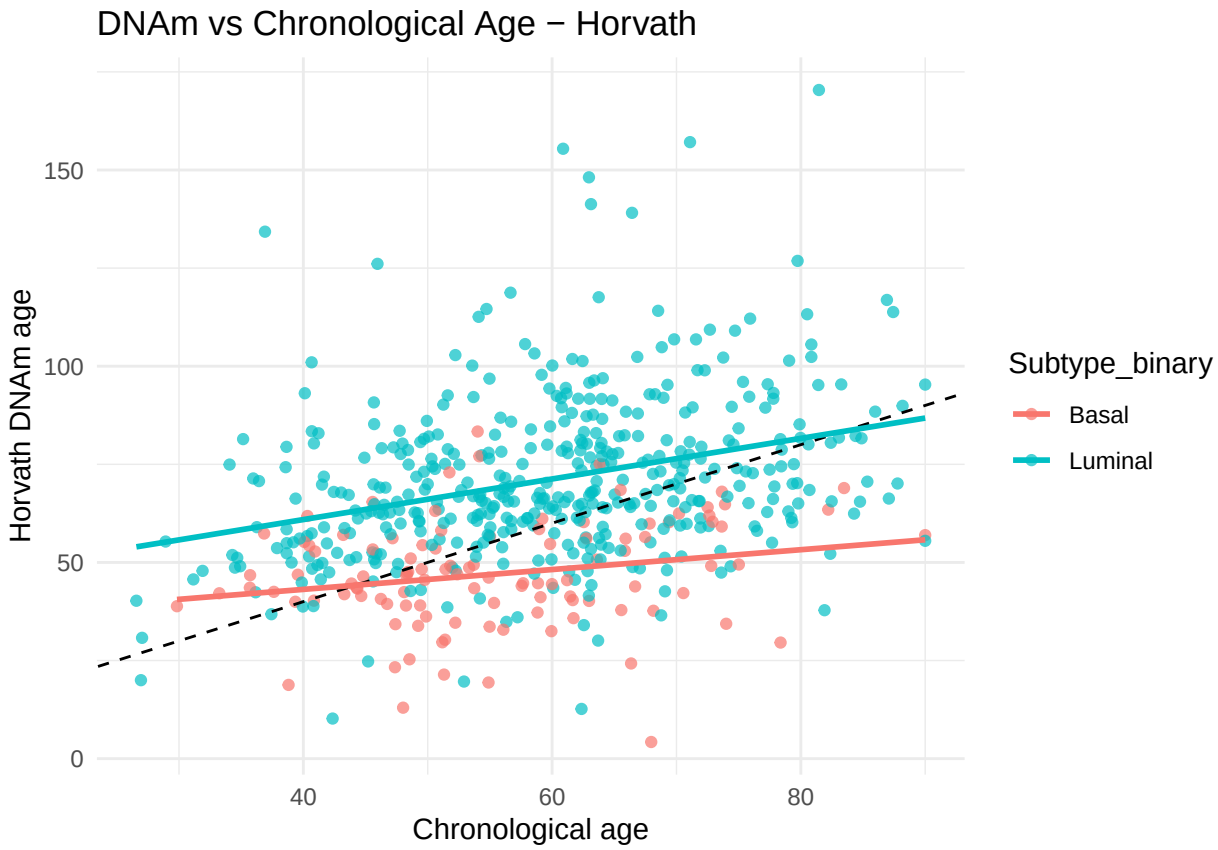
```
library(ggplot2)

for (clock in c("Horvath", "Hannum", "PhenoAge")) {
  p <- ggplot(delta_age_with_chrono_age, aes(x = chrono_age, y = .data[[clock]],
                                             colour = Subtype_binary)) +
    geom_point(alpha = 0.7) +
    geom_abline(slope = 1, linetype = "dashed") +
    geom_smooth(method = "lm", se = FALSE) +
    labs(x = "Chronological age", y = paste(clock, "DNAm age")) +
    ggtitle(paste("DNAm vs Chronological Age -", clock)) +
    theme_minimal()
  print(p)
}
```

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

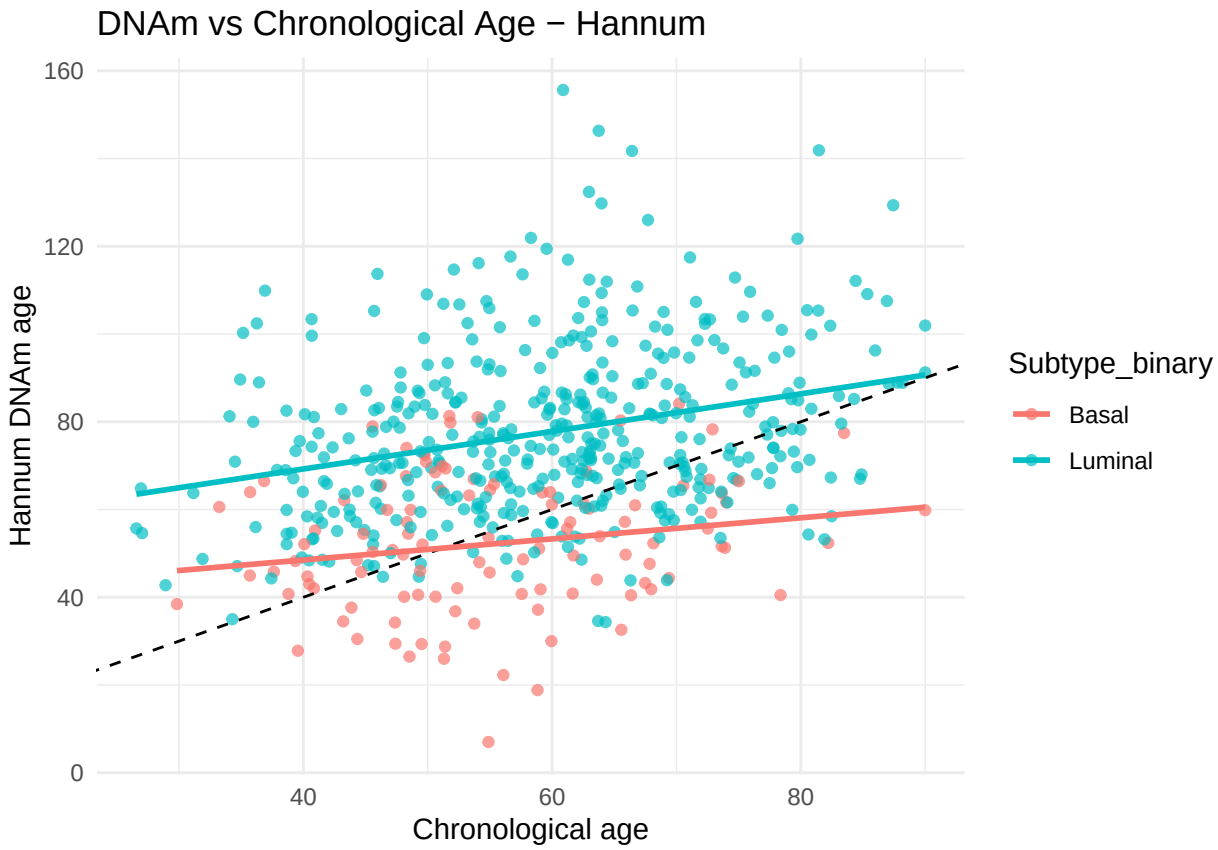
```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 9 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



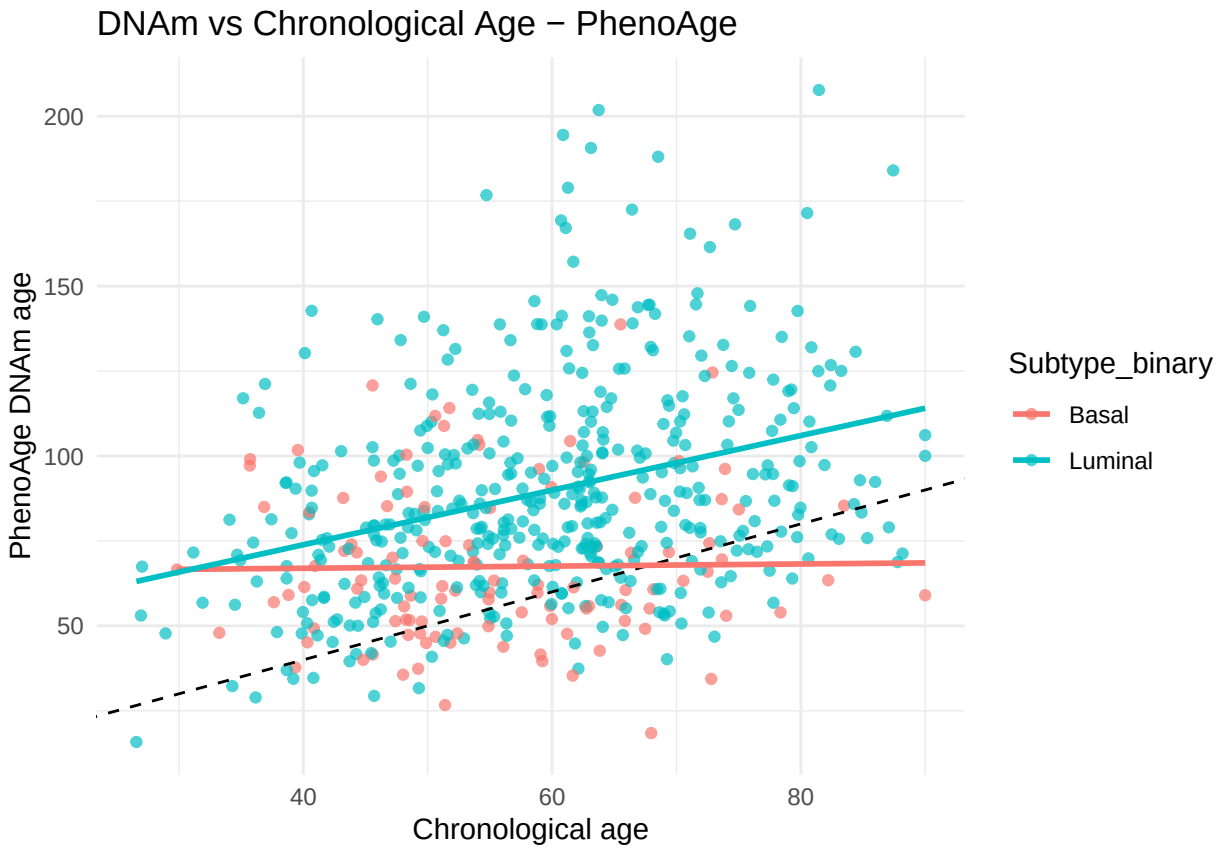

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range (`stat_smooth()`).
## Removed 9 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




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

```
## Warning: Removed 9 rows containing non-finite outside the scale range (`stat_smooth()`).
## Removed 9 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



#Interpretation Cross-clock consensus Luminal breast cancers: Epigenetic age > chronological age in all clocks (positive ΔAge) Age acceleration magnitude: PhenoAge > Horvath > Hannum Implies hormone-regulated, more-differentiated tumours accrue age-related methylation.

Basal breast cancers Epigenetic age or < chronological age (neutral or negative ΔAge) Weak slope DNAm age changes little with patient age Fits an undifferentiated, stem-like, often younger patient profile.

Why this matters Prognostic potential – ΔAge may help refine risk within each subtype. Biology – Subtype-specific ageing programs could influence tumour evolution and therapy response. Next analyses – Correlate ΔAge with immune infiltration, mutation burden, and survival to test mechanistic links.

Boxplots

```
library(ggplot2)
library(patchwork)

plot_list <- list()

for (clock in c("Horvath", "Hannum", "PhenoAge")) {
  delta_col <- paste0("delta_", clock)

  p <- ggplot(delta_age_with_chrono_age, aes(x = Subtype_binary, y = .data[[delta_col]],
                                             fill = Subtype_binary)) +
    geom_violin(trim = FALSE, alpha = 0.6) +
    geom_boxplot(width = 0.15, outlier.shape = NA) +
    labs(
      x = NULL,  # Remove x-axis label
```

```
      y = paste("Δ-age (", clock, ")", sep = ""),
      title = paste("Age_acce -", clock)
    ) +
    theme_minimal() +
    guides(fill = "none")

  plot_list[[clock]] <- p
}

# Combine plots and add a common x-axis label at the bottom
combined_plot <- (plot_list$Horvath + plot_list$Hannum + plot_list$PhenoAge) &
  theme(axis.title.x = element_blank())  # Avoid individual x-axis labels

# Add common x-label manually with patchwork annotation
(combined_plot +
  plot_annotation(
    caption = "Subtype_binary"
  )) &
  theme(
    plot.caption = element_text(hjust = 0.5, size = 12, face = "bold")
  )
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

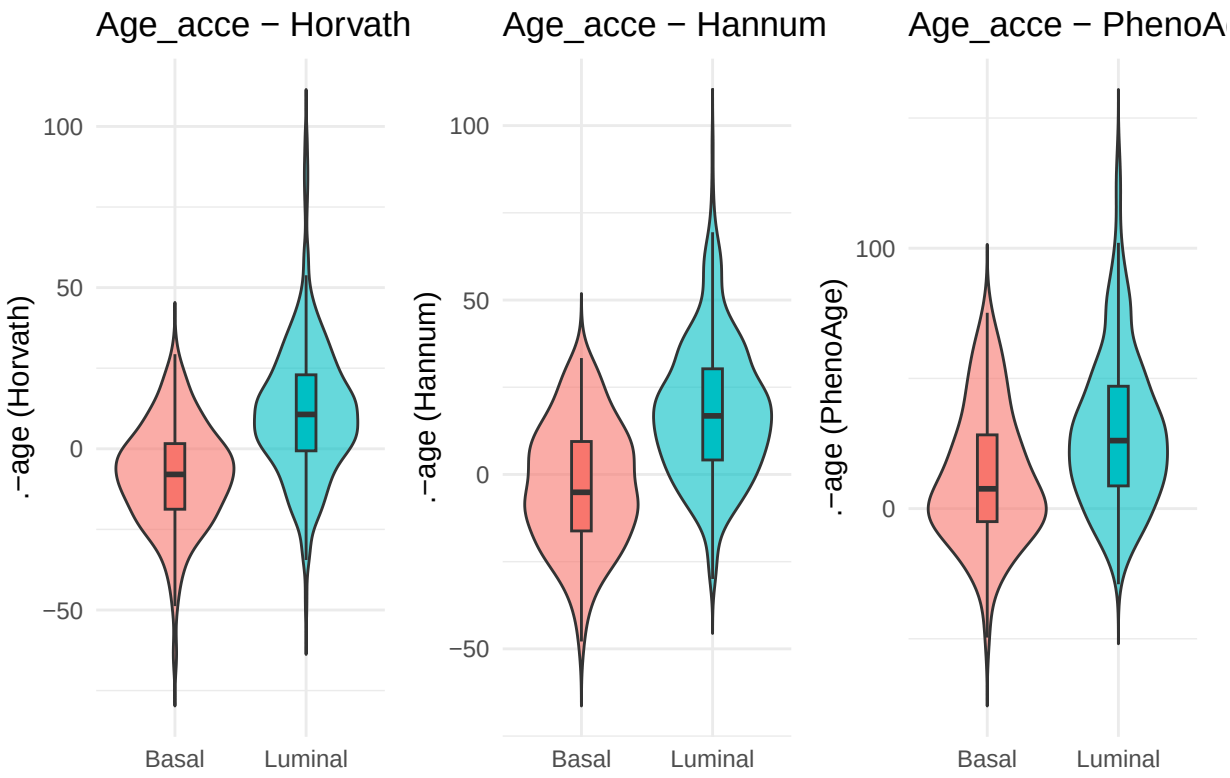
```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
library(ggplot2)
library(patchwork)

plot_list <- list()

for (clock in c("Horvath", "Hannum", "PhenoAge")) {
  delta_col <- paste0("delta_", clock)

  p <- ggplot(delta_age_with_chrono_age, aes(x = Subtype, y = .data[[delta_col]], fill = Subtype)) +
    geom_violin(trim = FALSE, alpha = 0.6) +
    geom_boxplot(width = 0.15, outlier.shape = NA) +
    labs(
      x = NULL,  # Remove x-axis label
      y = paste("Δ-age (", clock, ")", sep = ""),
      title = paste("Age_acce -", clock)
    ) +
    theme_minimal() +
    guides(fill = "none")

  plot_list[[clock]] <- p
}

# Combine plots and add a common x-axis label at the bottom
combined_plot <- (plot_list$Horvath + plot_list$Hannum + plot_list$PhenoAge) &
  theme(axis.title.x = element_blank())  # Avoid individual x-axis labels

# Add common x-label manually with patchwork annotation
```

```
combined_plot + plot_annotation(
  caption = "Subtype"
) & theme(plot.caption = element_text(hjust = 0.5, size = 12, face = "bold"))
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

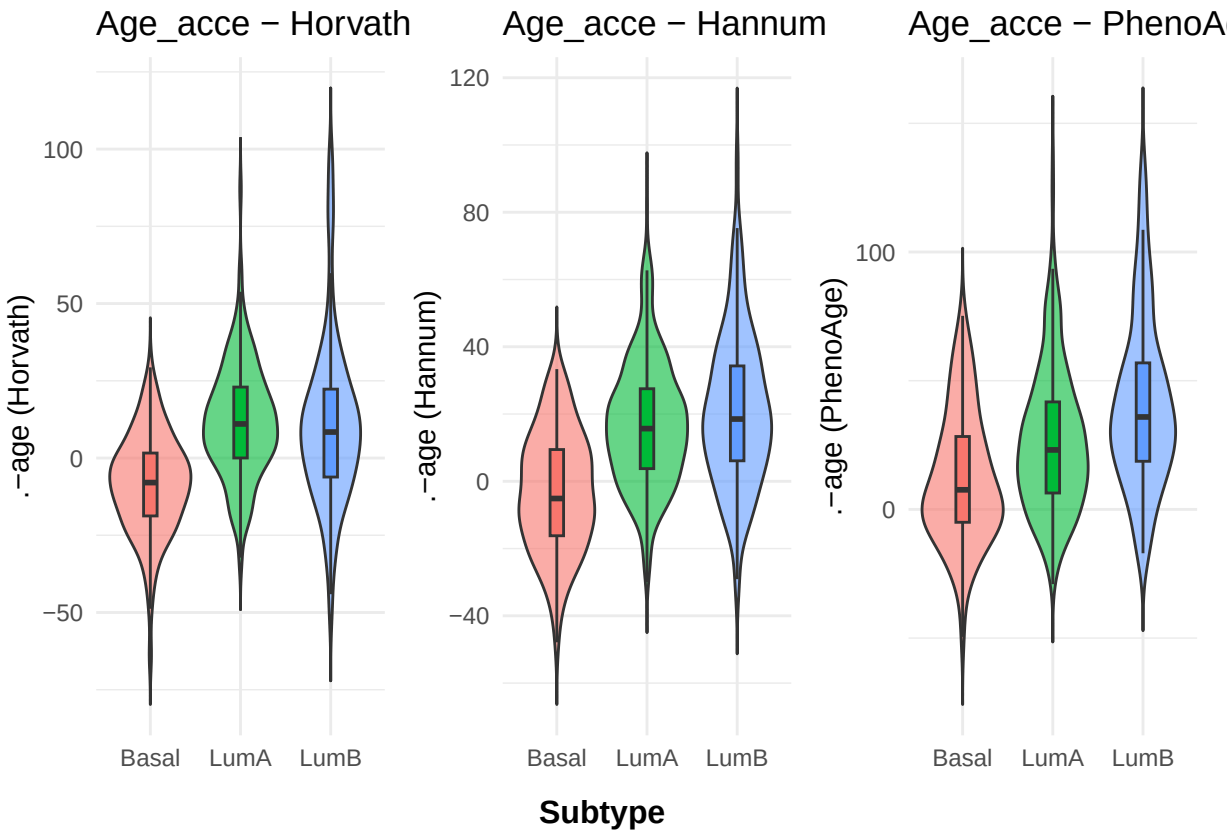
```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



Interpretation

In case of bi-class and tri class, Hannum clock shows higher Age accelaration than Hovarth and PhenoAge.

Across all three clocks, Luminal tumours display positive Δ-Age (epigenetic age acceleration).

Basal tumours cluster around zero or negative Δ-Age, indicating epigenetic age deceleration.

Confirms subtype-specific ageing biology: hormone-driven Luminals appear “older”, progenitor-like Basals “younger.”

Linear Model

```
### Adding Linear Model
for (clock in c("Horvath", "Hannum", "PhenoAge")) {
  delta_col <- paste0("delta_", clock)

  # Fit model
  fit_2 <- lm(formula = as.formula(
    paste0(delta_col, " ~ Subtype + age_at_index + ajcc_pathologic_stage")),
    data = clinical_data
  )

  cat("\n\n==== Linear Model Summary for Δ-age (", clock, ") ====\n")
  print(summary(fit_2))

}
```

```
##
##
## ==== Linear Model Summary for Δ-age ( Horvath ) ====
##
## Call:
## lm(formula = as.formula(paste0(delta_col, " ~ Subtype + age_at_index + ajcc_pathologic_stage")),
##     data = clinical_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.733 -11.565  -1.268   9.288  87.940
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      25.34575    9.09382   2.787  0.00551 **
## SubtypeLumA                      23.10380    2.14777  10.757  < 2e-16 ***
## SubtypeLumB                      21.30373    2.58559   8.239 1.37e-15 ***
## age_at_index                     -0.53511    0.06382  -8.385 4.63e-16 ***
## ajcc_pathologic_stageStage I     -4.23672    8.90371  -0.476  0.63439
## ajcc_pathologic_stageStage IA    -4.68397    8.87108  -0.528  0.59772
## ajcc_pathologic_stageStage IB     1.82009   20.65177   0.088  0.92980
## ajcc_pathologic_stageStage II    -1.56923   11.91419  -0.132  0.89526
## ajcc_pathologic_stageStage IIA   -3.70423    8.56898  -0.432  0.66571
## ajcc_pathologic_stageStage IIB   -4.11665    8.59277  -0.479  0.63208
## ajcc_pathologic_stageStage III    4.50122   20.72455   0.217  0.82814
## ajcc_pathologic_stageStage IIIA -6.44228    8.65946  -0.744  0.45723
## ajcc_pathologic_stageStage IIIB -8.63228   10.35680  -0.833  0.40495
## ajcc_pathologic_stageStage IIIC -7.95646    9.02616  -0.881  0.37845
## ajcc_pathologic_stageStage IV    -8.08844   11.40871  -0.709  0.47866
## ajcc_pathologic_stageStage X     -9.61543   11.91736  -0.807  0.42012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.82 on 529 degrees of freedom
##   (148 observations deleted due to missingness)
## Multiple R-squared:  0.2452, Adjusted R-squared:  0.2238
## F-statistic: 11.46 on 15 and 529 DF,  p-value: < 2.2e-16
```

```
##
##
##
## ==== Linear Model Summary for Δ-age ( Hannum ) ====
##
## Call:
## lm(formula = as.formula(paste0(delta_col, " ~ Subtype + age_at_index + ajcc_pathologic_stage")),
##     data = clinical_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.893 -12.437  -1.092  10.279  74.120
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      41.07048    8.63876   4.754 2.57e-06 ***
## SubtypeLumA                      23.60118    2.04029  11.568  < 2e-16 ***
## SubtypeLumB                      26.82711    2.45621  10.922  < 2e-16 ***
## age_at_index                     -0.60705    0.06062 -10.013  < 2e-16 ***
## ajcc_pathologic_stageStage I    -13.19880    8.45816  -1.560   0.1192
## ajcc_pathologic_stageStage IA    -9.59802    8.42716  -1.139   0.2552
## ajcc_pathologic_stageStage IB   -24.35566   19.61835  -1.241   0.2150
## ajcc_pathologic_stageStage II   -14.31835   11.31799  -1.265   0.2064
## ajcc_pathologic_stageStage IIA  -10.86190    8.14018  -1.334   0.1827
## ajcc_pathologic_stageStage IIB  -10.00284    8.16279  -1.225   0.2210
## ajcc_pathologic_stageStage III  -19.59887   19.68748  -0.995   0.3199
## ajcc_pathologic_stageStage IIIA -13.74897    8.22613  -1.671   0.0952 .
## ajcc_pathologic_stageStage IIIB -10.62146    9.83854  -1.080   0.2808
## ajcc_pathologic_stageStage IIIC -14.25191    8.57448  -1.662   0.0971 .
## ajcc_pathologic_stageStage IV   -11.85081   10.83781  -1.093   0.2747
## ajcc_pathologic_stageStage X    -12.19968   11.32101  -1.078   0.2817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.88 on 529 degrees of freedom
##   (148 observations deleted due to missingness)
## Multiple R-squared:  0.3071, Adjusted R-squared:  0.2874
## F-statistic: 15.63 on 15 and 529 DF,  p-value: < 2.2e-16
##
##
##
## ==== Linear Model Summary for Δ-age ( PhenoAge ) ====
##
## Call:
## lm(formula = as.formula(paste0(delta_col, " ~ Subtype + age_at_index + ajcc_pathologic_stage")),
##     data = clinical_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.866 -19.437  -4.609  15.627 115.008
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      39.39702   13.75327   2.865  0.00434 **
```

```
## SubtypeLumA                       16.03634    3.24824   4.937 1.07e-06 ***
## SubtypeLumB                       29.96150    3.91039   7.662 8.79e-14 ***
## age_at_index                      -0.30426    0.09652  -3.152  0.00171 **
## ajcc_pathologic_stageStage I     -13.43942   13.46575  -0.998  0.31871
## ajcc_pathologic_stageStage IA    -10.73882   13.41640  -0.800  0.42382
## ajcc_pathologic_stageStage IB     -9.18008   31.23324  -0.294  0.76893
## ajcc_pathologic_stageStage II      0.93705   18.01873   0.052  0.95855
## ajcc_pathologic_stageStage IIA   -10.11878   12.95951  -0.781  0.43527
## ajcc_pathologic_stageStage IIB   -13.18628   12.99550  -1.015  0.31072
## ajcc_pathologic_stageStage III    11.54877   31.34330   0.368  0.71268
## ajcc_pathologic_stageStage IIIA   -8.21745   13.09635  -0.627  0.53063
## ajcc_pathologic_stageStage IIIB -23.92369   15.66338  -1.527  0.12727
## ajcc_pathologic_stageStage IIIC -12.49788   13.65094  -0.916  0.36033
## ajcc_pathologic_stageStage IV     -5.69594   17.25425  -0.330  0.74144
## ajcc_pathologic_stageStage X      -0.95697   18.02353  -0.053  0.95768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.46 on 529 degrees of freedom
##   (148 observations deleted due to missingness)
## Multiple R-squared:  0.1235, Adjusted R-squared:  0.09864
## F-statistic: 4.969 on 15 and 529 DF,  p-value: 3.526e-09
```

Interpretation & Implications

Subtype Effect: All three epigenetic clocks detect significantly higher Δ-age in Luminal subtypes, indicating accelerated biological aging in these patients.

Age Effect: Older patients tend to have lower Δ-age, suggesting that epigenetic age plateaus or compresses relative to chronological age in older individuals.

Stage Effect: Surprisingly, tumor stage does not significantly predict Δ-age in any model. This might indicate that biological age acceleration is more subtype-driven than progression-driven.

Model Strength: Hannum has the best overall fit (R^2 0.30), followed by Horvath, and PhenoAge lags significantly behind.

PhenoAge may reflect more systemic, metabolic, or immune-related aging not captured in this model.