MECP2 Analysis

Andrew Willems and Tian Hong

5/23/2022

Objective

We are doing analysis on the new set of cohorts from the Krishnan lab for MECP2. First, we use intraclass correlation coefficient (ICC) values of the Cohort, Cell type, Cell number, and Image variables to determine if we need to build linear mixed effects models. After investigating if we need LMEs we then create heat maps of the MECP2 data. We compare the differences in means between the various conditions.

Step One

Load needed packages. `effectsize` is used to calculate the effect sizes of the differences in our various conditions/treatments. `ggpubr` is for the grouped plot support. `ggsignif` is used to add statistical results to ggplot plots. `gt` is used for making the nice tables. `ICC` is used to calculate the intraclass correlation coefficient to tell us if we can treat our predictors (variables) as independent or not. `magrittr` is a package that allows us to use pipes (%>%) in our code. `nlme` is the package that performs the linear mixed effects (lme) model fits. `rstatix` is used to do the pairwise t-tests and p-value correction. `tidyverse` is used for data manipulation. `webshot` is used to save our gt tables as `.png` files.

Step Two

Load the data and make separate data frames that are comprised of only 6 or 12 week data. The warning here is okay. When I make all columns numeric it introduces some `NAs` because not all columns have the same number of rows (some just have no data in that row and therefore they get an `NA`). When calculating the mean later those rows with `NAs` are not included in the calculation.

Step Three: Getting straight mean for all of our data

Six week old data

Twelve week old data

Step Four: Adding the means to our overall data frames

Filtering to just naive `Condition`

```
mecp2_data_processed_6wk <- mecp2_data_processed_6wk %>% filter(Condition=="NW" | Condition=="NH")
mecp2_data_processed_12wk <- mecp2_data_processed_12wk %>% filter(Condition=="NW" | Condition=="NH")
```

Relabeling `NW` and `NH` as `WT` and `Het` respectively

```
mecp2_data_processed_6wk$Condition <- gsub(x = mecp2_data_processed_6wk$Condition, pattern = "NW", repla
mecp2_data_processed_6wk$Condition <- gsub(x = mecp2_data_processed_6wk$Condition, pattern = "NH", repla
```

```
mecp2_data_processed_12wk$Condition <- gsub(x = mecp2_data_processed_12wk$Condition, pattern = "NW", re
mecp2_data_processed_12wk$Condition <- gsub(x = mecp2_data_processed_12wk$Condition, pattern = "NH", re
```

Now making the hemisphere all the same (LH) so that our analysis is correct for means

```
mecp2_data_processed_6wk$Hemisphere <- gsub(x = mecp2_data_processed_6wk$Hemisphere, pattern = "RH", re
```

```
mecp2_data_processed_12wk$Hemisphere <- gsub(x = mecp2_data_processed_12wk$Hemisphere, pattern = "RH",
```

Checking to see if my means are the same as Logan's and Tian's

```
by_cohort_age_type_6wk <- mecp2_data_processed_6wk %>% group_by(Cohort, Time,
                                                                Cell_type,
                                                                Condition,
                                                                Hemisphere)
by_cohort_age_type_6wk
```

```
## # A tibble: 184 x 8
## # Groups:   Cohort, Time, Cell_type, Condition, Hemisphere [12]
##    Cell_type Cohort  Condition Hemisphere Image Cell_number Intensity Time
##    <chr>     <chr>   <chr>     <chr>      <chr> <chr>           <dbl> <chr>
##  1 PNN-neg   #102319 WT        LH         1     1               1536. 6 wk
##  2 PNN-neg   #102319 WT        LH         1     2               1163. 6 wk
##  3 PNN-neg   #102319 WT        LH         1     3               1702. 6 wk
##  4 PNN-neg   #102319 WT        LH         1     4               1570. 6 wk
##  5 PNN-neg   #102319 WT        LH         2     1               1316. 6 wk
##  6 PNN-neg   #102319 WT        LH         2     2                960. 6 wk
##  7 PNN-neg   #102319 WT        LH         2     3               1202. 6 wk
##  8 PNN-neg   #102319 WT        LH         2     4               1405. 6 wk
##  9 PNN-neg   #103119 WT        LH         1     1                784. 6 wk
## 10 PNN-neg   #103119 WT        LH         1     2                845. 6 wk
## # ... with 174 more rows
```

```
## `summarise()` has grouped output by 'Cohort', 'Time', 'Cell_type', 'Condition'.
## You can override using the `.groups` argument.
```

```
## `summarise()` has grouped output by 'Cohort', 'Time', 'Cell_type', 'Condition'.
## You can override using the `.groups` argument.
```

Filtering to just PNN-pos or PNN-neg cell types

```
mecp2_6_pos <- by_cohort_age_type_6wk %>% filter(Cell_type=="PNN-pos")
mecp2_6_neg <- by_cohort_age_type_6wk %>% filter(Cell_type=="PNN-neg")

mecp2_12_pos <- by_cohort_age_type_12wk %>% filter(Cell_type=="PNN-pos")
mecp2_12_neg <- by_cohort_age_type_12wk %>% filter(Cell_type=="PNN-neg")
```

```
## # A tibble: 184 x 8
## # Groups:   Cohort, Time, Cell_type, Condition, Hemisphere [12]
##    Cell_type Cohort  Condition Hemisphere Image Cell_number Intensity Time
##    <chr>     <chr>   <chr>     <chr>      <chr> <chr>           <dbl> <chr>
##  1 PNN-pos   #102319 WT        LH         1     1               3551. 6 wk
##  2 PNN-pos   #102319 WT        LH         1     2               2704. 6 wk
##  3 PNN-pos   #102319 WT        LH         1     3               4440. 6 wk
##  4 PNN-pos   #102319 WT        LH         2     1               4044. 6 wk
##  5 PNN-pos   #102319 WT        LH         2     2               3149. 6 wk
##  6 PNN-pos   #102319 WT        LH         2     3               3277. 6 wk
##  7 PNN-pos   #103119 WT        LH         1     1               2329. 6 wk
##  8 PNN-pos   #103119 WT        LH         1     2               2193. 6 wk
##  9 PNN-pos   #103119 WT        LH         1     3               2579. 6 wk
## 10 PNN-pos   #103119 WT        LH         1     4               2166. 6 wk
```

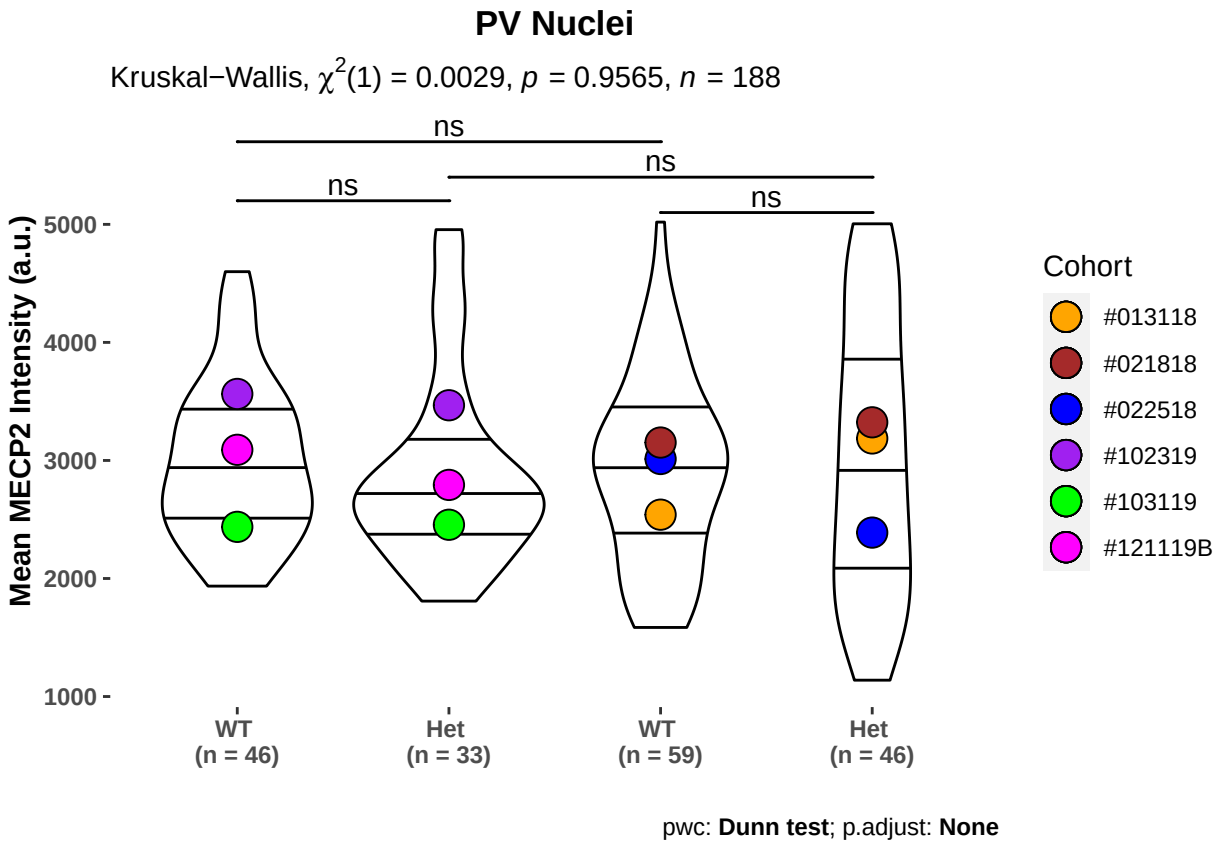
```
## # ... with 174 more rows
```

The table that contains the means that are equivalent to Logan's

```
## # A tibble: 223 x 8
## # Groups:   Cohort, Time, Cell_type, Condition, Hemisphere [12]
##    Cell_type Cohort  Condition Hemisphere Image Cell_number Intensity Time
##    <chr>     <chr>   <chr>     <chr>      <chr> <chr>           <dbl> <chr>
##  1 PNN-neg   #102319 WT        LH         1     1               1536. 6 wk
##  2 PNN-neg   #102319 WT        LH         1     2               1163. 6 wk
##  3 PNN-neg   #102319 WT        LH         1     3               1702. 6 wk
##  4 PNN-neg   #102319 WT        LH         1     4               1570. 6 wk
##  5 PNN-neg   #102319 WT        LH         2     1               1316. 6 wk
##  6 PNN-neg   #102319 WT        LH         2     2                960. 6 wk
##  7 PNN-neg   #102319 WT        LH         2     3               1202. 6 wk
##  8 PNN-neg   #102319 WT        LH         2     4               1405. 6 wk
##  9 PNN-neg   #103119 WT        LH         1     1                784. 6 wk
## 10 PNN-neg   #103119 WT        LH         1     2                845. 6 wk
## # ... with 213 more rows
```

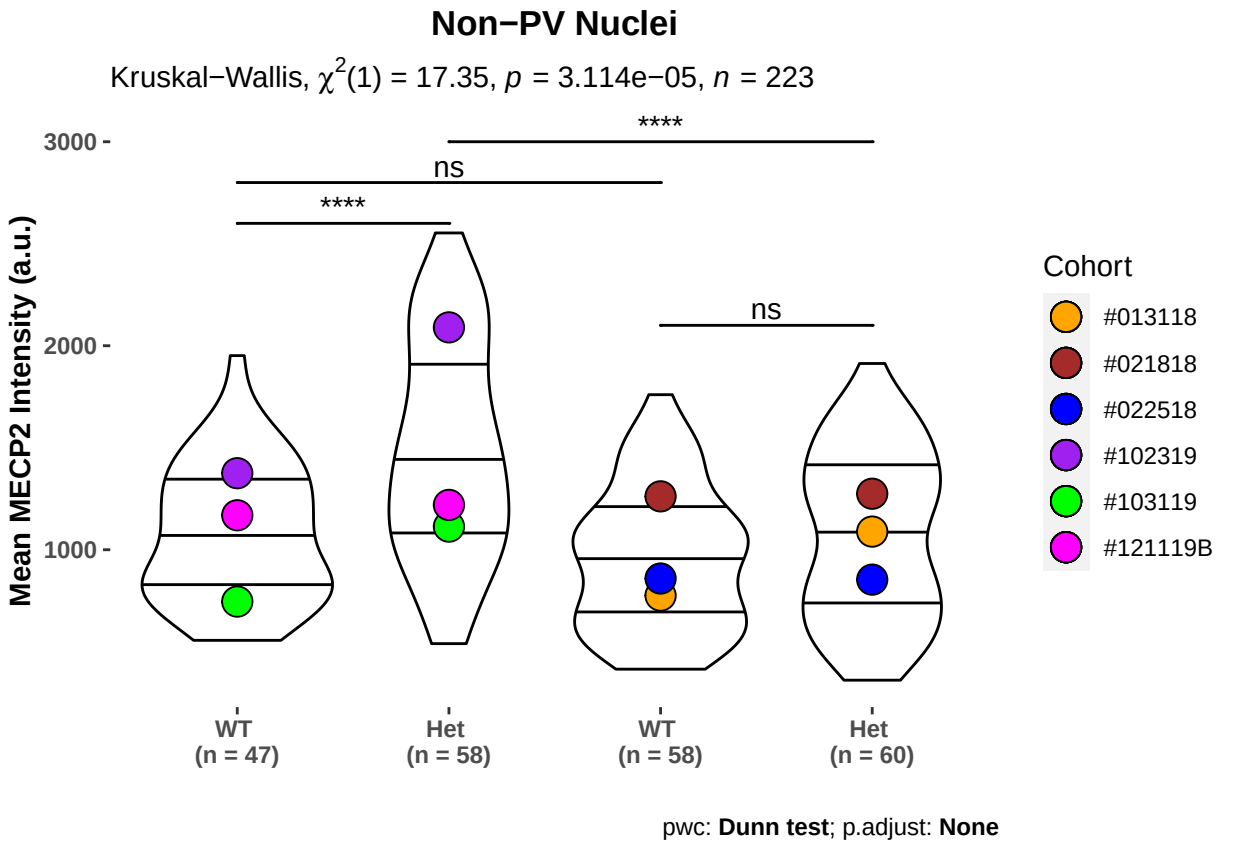
```
## # A tibble: 184 x 8
## # Groups:   Cohort, Time, Cell_type, Condition, Hemisphere [12]
##    Cell_type Cohort  Condition Hemisphere Image Cell_number Intensity Time
##    <chr>     <fct>   <chr>     <chr>      <chr> <chr>           <dbl> <fct>
##  1 PNN-pos   #102319 WT        LH         1     1               3551. 6 wk
##  2 PNN-pos   #102319 WT        LH         1     2               2704. 6 wk
##  3 PNN-pos   #102319 WT        LH         1     3               4440. 6 wk
##  4 PNN-pos   #102319 WT        LH         2     1               4044. 6 wk
##  5 PNN-pos   #102319 WT        LH         2     2               3149. 6 wk
##  6 PNN-pos   #102319 WT        LH         2     3               3277. 6 wk
##  7 PNN-pos   #103119 WT        LH         1     1               2329. 6 wk
##  8 PNN-pos   #103119 WT        LH         1     2               2193. 6 wk
##  9 PNN-pos   #103119 WT        LH         1     3               2579. 6 wk
## 10 PNN-pos   #103119 WT        LH         1     4               2166. 6 wk
## # ... with 174 more rows
```

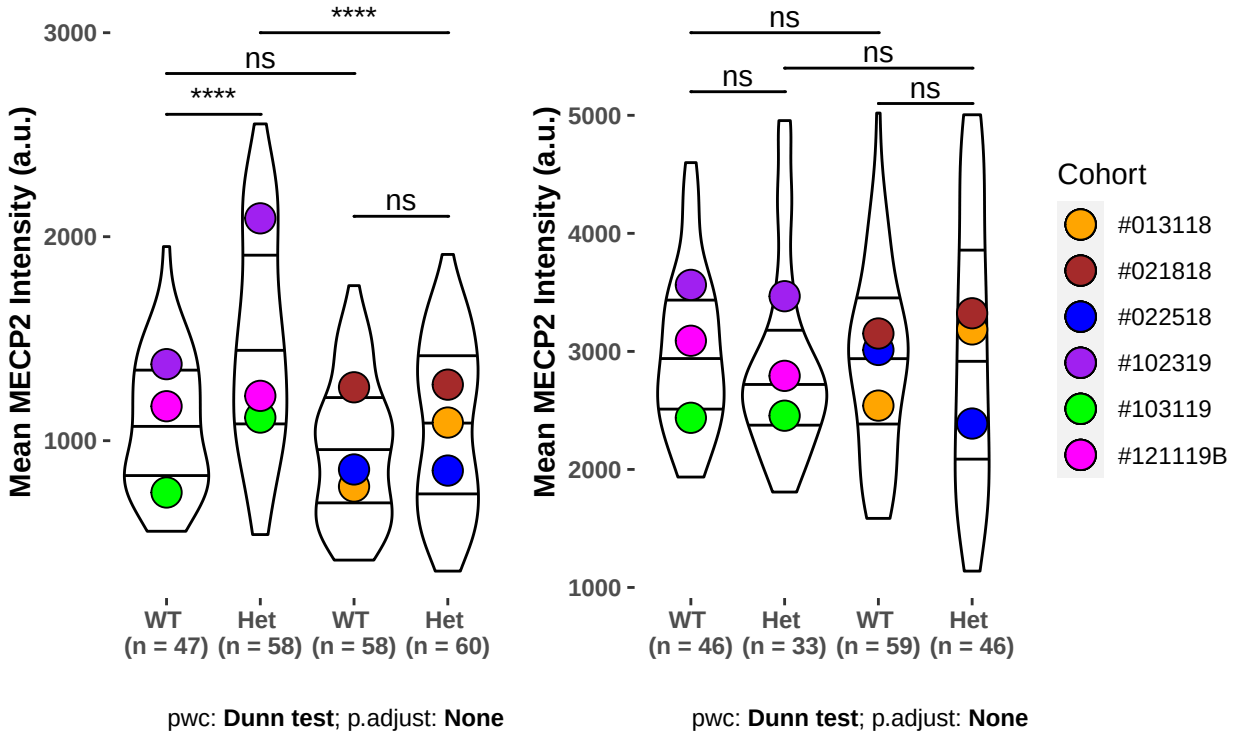
Now doing all the statistical analysis and plotting for the PV Nuclei (`PNN-pos`) containing samples



Now doing all the statistical analysis and plotting for the Non-PV Nuclei (`PNN-neg`) containing samples

```
total_plot_neg
```


Non−PV Nuclei
Kruskal−Wallis, $\chi^2(1) = 17.35$, $p = 3.114e{-}05$, $n = 223$
Mean MECP2 Intensity (a.u.)
3000
2000
1000
****
ns
****
ns
WT
(n = 47)
Het
(n = 58)
WT
(n = 58)
Het
(n = 60)
Cohort
#013118
#021818
#022518
#102319
#103119
#121119B
pwc: Dunn test; p.adjust: None


e
Non−PV Nuclei
Kruskal−Wallis, $\chi^2(1) = 17.35$, p
Mean MECP2 Intensity (a.u.)
3000
2000
1000
****
ns
****
ns
WT
(n = 47)
Het
(n = 58)
WT
(n = 58)
Het
(n = 60)
pwc: Dunn test; p.adjust: None
f
PV Nuclei
Kruskal−Wallis, $\chi^2(1) = 0.0029$, $p = 0.9565$, n =
Mean MECP2 Intensity (a.u.)
5000
4000
3000
2000
1000
ns
ns
ns
ns
WT
(n = 46)
Het
(n = 33)
WT
(n = 59)
Het
(n = 46)
Cohort
#013118
#021818
#022518
#102319
#103119
#121119B
pwc: Dunn test; p.adjust: None

Now performing ICC analysis on the combinations we have previously tested to see if any of the variables have high levels of dependence

ICC for Non-PV MECP2 Data

Intraclass Correlation Coefficient (ICC) for Mean 6 and 12 week Non-PV MECP2 data.

Cohort	Cell number	Image
0.4913868	-0.01769006	-0.009018774

ICC for PV MECP2 Data

Intraclass Correlation Coefficient (ICC) for Mean 6 and 12 week PV MECP2 data.

Cohort	Cell number	Image
0.1473937	-0.010394	-0.00929889

Building the lme for non-PV nuclei because of high ICC for `Cohort`

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: mecp2_6_12_neg
##        AIC      BIC    logLik
##   3262.944 3276.573 -1627.472
##
## Random effects:
##  Formula: ~1 | Cohort
##         (Intercept) Residual
## StdDev:    271.9573 342.4689
##
## Fixed effects:  Intensity ~ Time
##                 Value Std.Error  DF   t-value p-value
## (Intercept) 1302.0050  161.2702 217  8.073437  0.0000
## Time12 wk   -283.2792  227.7900   4 -1.243598  0.2816
##  Correlation:
##           (Intr)
## Time12 wk -0.708
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.3131246 -0.7414420 -0.1484831  0.7516162  2.3371940
##
## Number of Observations: 223
## Number of Groups: 6
```

Doing 6 week WT vs. Het lme model

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: six_wk_comp_df
##        AIC      BIC    logLik
##   1502.783 1513.399 -747.3915
##
## Random effects:
##  Formula: ~1 | Cohort
##         (Intercept) Residual
## StdDev:    342.4733 282.1681
##
```

```
## Fixed effects:  Intensity ~ Condition
##                Value Std.Error  DF   t-value p-value
## (Intercept) 1476.991 203.11322 101  7.271763       0
## ConditionWT -392.603  56.03099 101 -7.006891       0
##  Correlation:
##             (Intr)
## ConditionWT -0.123
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.9108768 -0.5592748  0.1377171  0.6919680  2.1787698
##
## Number of Observations: 105
## Number of Groups: 3
```

Doing 12 week WT vs. Het lme model

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: twelve_week_comp
##       AIC      BIC    logLik
##   1723.07 1734.152 -857.5349
##
## Random effects:
##  Formula: ~1 | Cohort
##         (Intercept) Residual
## StdDev:    169.2587  336.254
##
## Fixed effects:  Intensity ~ Condition
##                 Value Std.Error  DF   t-value p-value
## (Intercept) 1072.2688 107.84751 114  9.942453  0.0000
## ConditionWT -109.4505  62.46691 114 -1.752135  0.0824
##  Correlation:
##             (Intr)
## ConditionWT -0.285
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.1138936 -0.7085800 -0.1759541  0.7399168  2.1926509
##
## Number of Observations: 118
## Number of Groups: 3
```

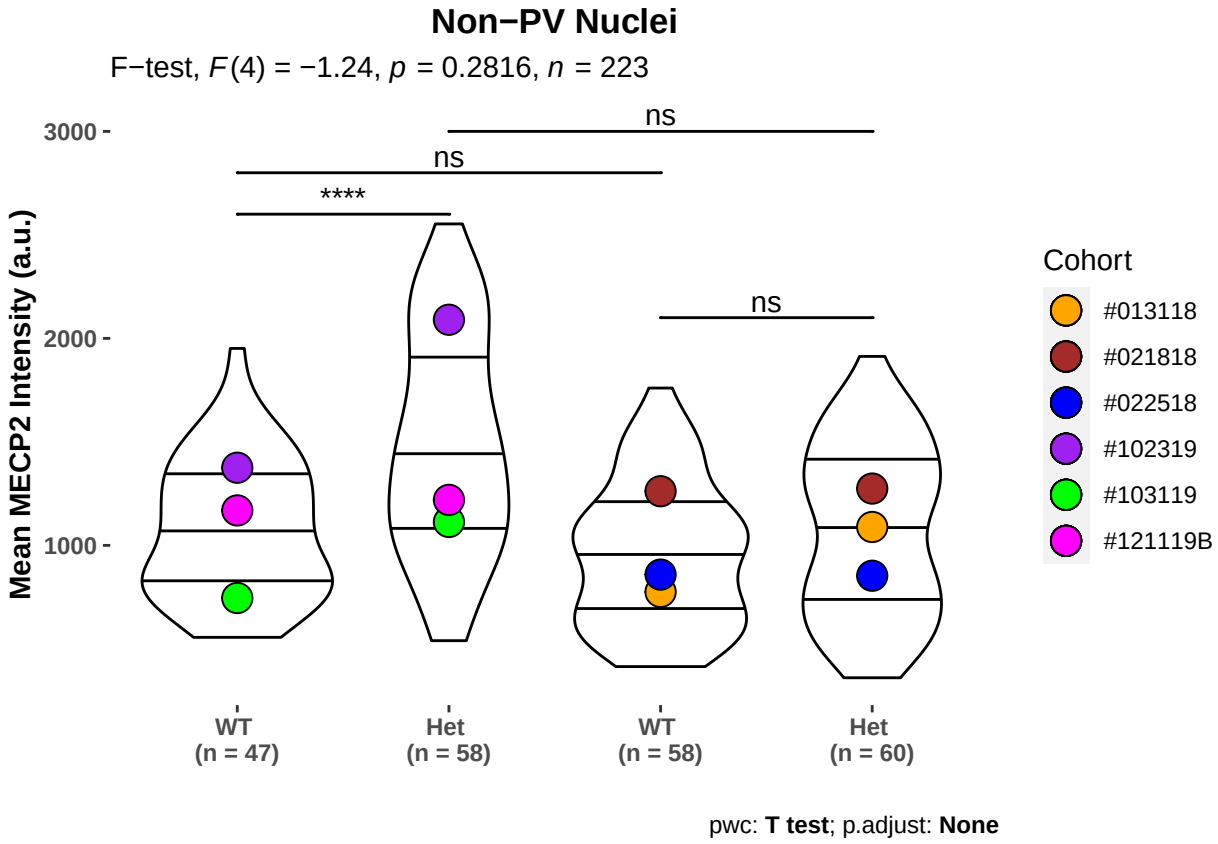
Doing 6 week WT vs. 12 week WT lme model

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: wt_v_wt_comp
##        AIC      BIC    logLik
##   1478.003 1488.619 -735.0014
##
## Random effects:
##  Formula: ~1 | Cohort
##         (Intercept) Residual
## StdDev:    231.3728 244.9482
##
## Fixed effects:  Intensity ~ Time
##                 Value Std.Error DF   t-value p-value
```

```
## (Intercept) 1095.4231  139.6938 99  7.841603  0.0000
## Time12 wk   -130.7135  196.8815  4 -0.663920  0.5431
##  Correlation:
##           (Intr)
## Time12 wk -0.71
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -1.6188978 -0.6851402 -0.1259129  0.5716026  2.6516168
##
## Number of Observations: 105
## Number of Groups: 6
```

Doing 6 week Het vs. 12 week Het lme model

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: het_v_het_comp
##        AIC     BIC    logLik
##   1734.077 1745.16 -863.0385
##
## Random effects:
##  Formula: ~1 | Cohort
##         (Intercept) Residual
## StdDev:    323.7375  336.934
##
## Fixed effects:  Intensity ~ Time
##                 Value Std.Error  DF   t-value p-value
## (Intercept) 1474.8801  193.7262 112  7.613220  0.0000
## Time12 wk   -402.6113  273.8470   4 -1.470205  0.2155
##  Correlation:
##           (Intr)
## Time12 wk -0.707
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.05917903 -0.72272132  0.06214897  0.72591040  2.12980364
##
## Number of Observations: 118
## Number of Groups: 6
```



Getting an lme of PV nuclei even though their ICC is low (0.14) and plotting to see if anything changes

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: mecp2_6_12_pos
##        AIC      BIC    logLik
##   3003.947 3016.807 -1497.974
##
## Random effects:
##  Formula: ~1 | Cohort
##         (Intercept) Residual
## StdDev:    306.3937 808.2376
##
## Fixed effects:  Intensity ~ Time
##                 Value Std.Error  DF   t-value p-value
## (Intercept) 2971.0410  200.1857 178 14.841423  0.0000
## Time12 wk    -20.6061  279.3225   4 -0.073772  0.9447
##  Correlation:
##           (Intr)
## Time12 wk -0.717
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.0670725 -0.6178736 -0.1014825  0.5883319  2.5820325
##
## Number of Observations: 184
## Number of Groups: 6
```

Doing 6 week WT vs. Het lme model

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: six_wk_comp_df
##        AIC      BIC    logLik
##   1241.236 1250.713 -616.6178
##
## Random effects:
##  Formula: ~1 | Cohort
##         (Intercept) Residual
## StdDev:    429.1473 563.1004
##
## Fixed effects:  Intensity ~ Condition
##                Value Std.Error DF   t-value p-value
## (Intercept) 2909.863   269.932 75 10.779983  0.0000
## ConditionWT  111.451   130.170 75  0.856196  0.3946
##  Correlation:
##             (Intr)
## ConditionWT -0.28
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.86984841 -0.60985792 -0.09029544  0.55439234  2.71860368
##
## Number of Observations: 79
## Number of Groups: 3
```

Doing 12 week WT vs. Het lme model

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: twelve_week_comp
##        AIC      BIC    logLik
##   1746.603 1757.219 -869.3014
##
## Random effects:
##  Formula: ~1 | Cohort
##         (Intercept) Residual
## StdDev:    117.6614 947.6272
##
## Fixed effects:  Intensity ~ Condition
##                 Value Std.Error  DF  t-value p-value
## (Intercept) 3022.8126  157.0944 101 19.24201  0.0000
## ConditionWT -123.8546  188.4236 101 -0.65732  0.5125
##  Correlation:
##             (Intr)
## ConditionWT -0.675
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -1.93025632 -0.80769522 -0.02099252  0.65800638  2.13376425
##
## Number of Observations: 105
## Number of Groups: 3
```

Doing 6 week WT vs. 12 week WT lme model

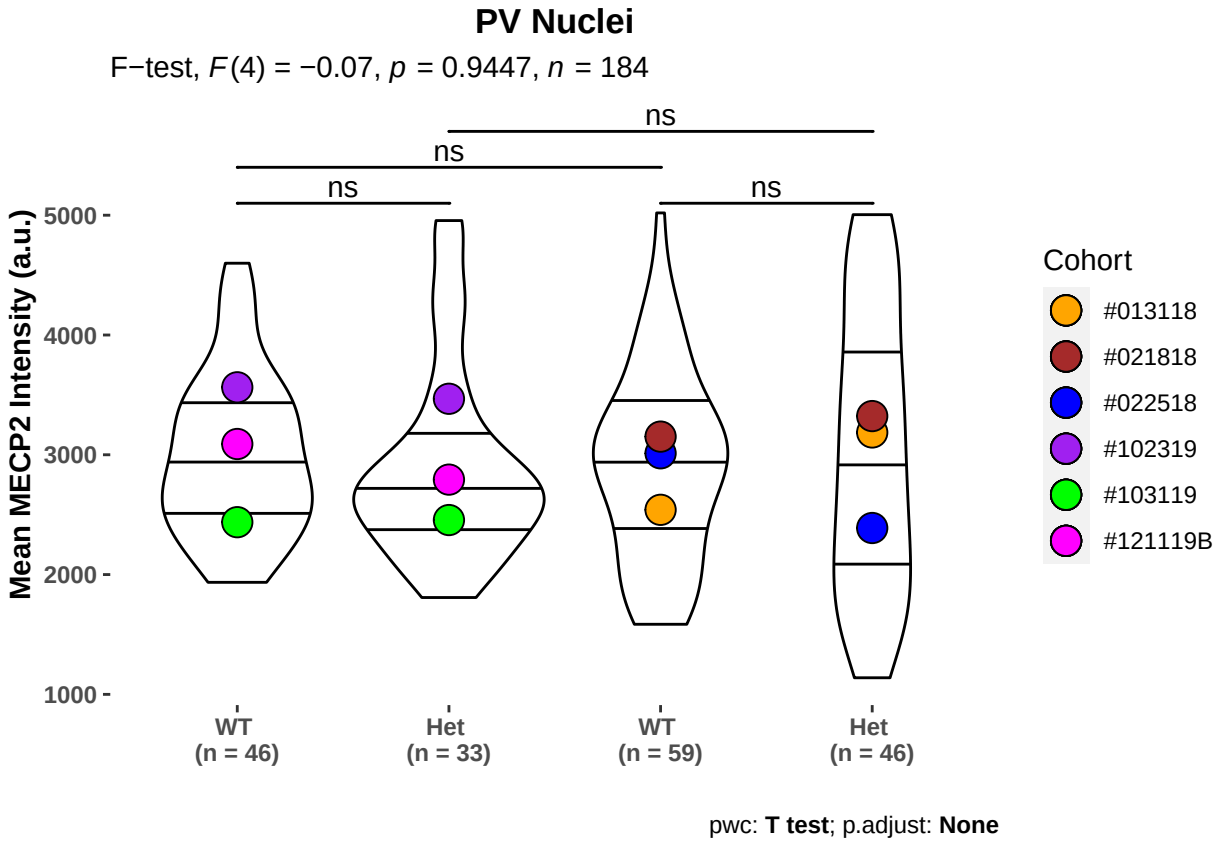
```
## Linear mixed-effects model fit by maximum likelihood
##   Data: wt_v_wt_comp
```

```
##        AIC      BIC    logLik
##   1675.553 1686.169 -833.7766
##
## Random effects:
##  Formula: ~1 | Cohort
##         (Intercept) Residual
## StdDev:    341.3554 646.3952
##
## Fixed effects:  Intensity ~ Time
##                Value Std.Error DF   t-value p-value
## (Intercept) 3023.112  221.3343 99 13.658576   0.000
## Time12 wk   -122.242  309.5273  4 -0.394931   0.713
##  Correlation:
##           (Intr)
## Time12 wk -0.715
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.12414530 -0.61017607 -0.06594409  0.48810881  2.95068401
##
## Number of Observations: 105
## Number of Groups: 6
```

Doing 6 week Het vs. 12 week Het lme model

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: het_v_het_comp
##        AIC      BIC    logLik
##   1322.429 1331.907 -657.2145
##
## Random effects:
##  Formula: ~1 | Cohort
##         (Intercept) Residual
## StdDev:    308.5803 960.9645
##
## Fixed effects:  Intensity ~ Time
##                 Value Std.Error DF   t-value p-value
## (Intercept) 2900.7562  247.7088 73 11.710347     0.0
## Time12 wk     91.7544  338.8888  4  0.270751     0.8
##  Correlation:
##           (Intr)
## Time12 wk -0.731
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -1.6857894 -0.6249335 -0.1062797  0.5139219  1.9003288
##
## Number of Observations: 79
## Number of Groups: 6
```

PV Nuclei lme plot



Now doing ICC for just the non-pv het samples between 6 and 12 weeks to see if the ICC is large for this specific comparison or not

ICC for Non-PV Het Only MECP2 Data

Intraclass Correlation Coefficient (ICC) for Mean 6 and 12 week Non-PV Het Only MECP2 data.

Cohort	Cell number	Image
0.6108359	-0.03353344	-0.01574835

Given the change from very statistically significant to non-significance between the 6 week and 12 week `Het` groups I wanted to see how many correlated neurons equaled one uncorrelated neuron and how many more neurons we would need to see a difference between these groups. I took the total number of neurons from the MECP2 negative group and divided it by the number of cohorts (because `cohort` is the variable with a high ICC). From this I got the average cluster size `M`. From there I calculated the Design Effect (`deff`). This tells us how many dependent neurons equal one uncorrelated neuron. From this we can get the effective sample size (`neff`) which tells us the equivalent number of cohorts if there was no correlation/clustering.

Our results show that about 11.58 cohorts is what we would need to get a sample size that would be equivalent to a sample size that had no correlation/clustering. This is about 1.93 times as many cohorts. (e.g. 12 needed instead of the 6 currently done). Given this we recommend an additional `n` of 6 mouse cohorts for analysis.

Power Analysis for Non-PV All Samples

Metrics to determine what sample size is needed to determine if their are statistically significant differences that are independent of cohort

M	Design effect	Effective size
37.17	19.26	11.58

Non-PV Het only recommendation

Power Analysis for Non-PV Het Only Samples

Metrics to determine what sample size is needed to determine if their are statistically significant differences that are independent of cohort

M	Design effect	Effective size
19.67	13.01	9.07

Checking if WT only non-PV has high ICC for Keerthi

ICC for Non-PV WT Only MECP2 Data

Intraclass Correlation Coefficient (ICC) for Mean 6 and 12 week Non-PV WT Only MECP2 data.

Cohort	Cell number	Image
0.5389673	0.003930653	-0.01774983

Non-PV WT only recommendation

Power Analysis for Non-PV WT Only Samples

Metrics to determine what sample size is needed to determine if their are statistically significant differences that are independent of cohort

M	Design effect	Effective size
17.5	10.43	10.07

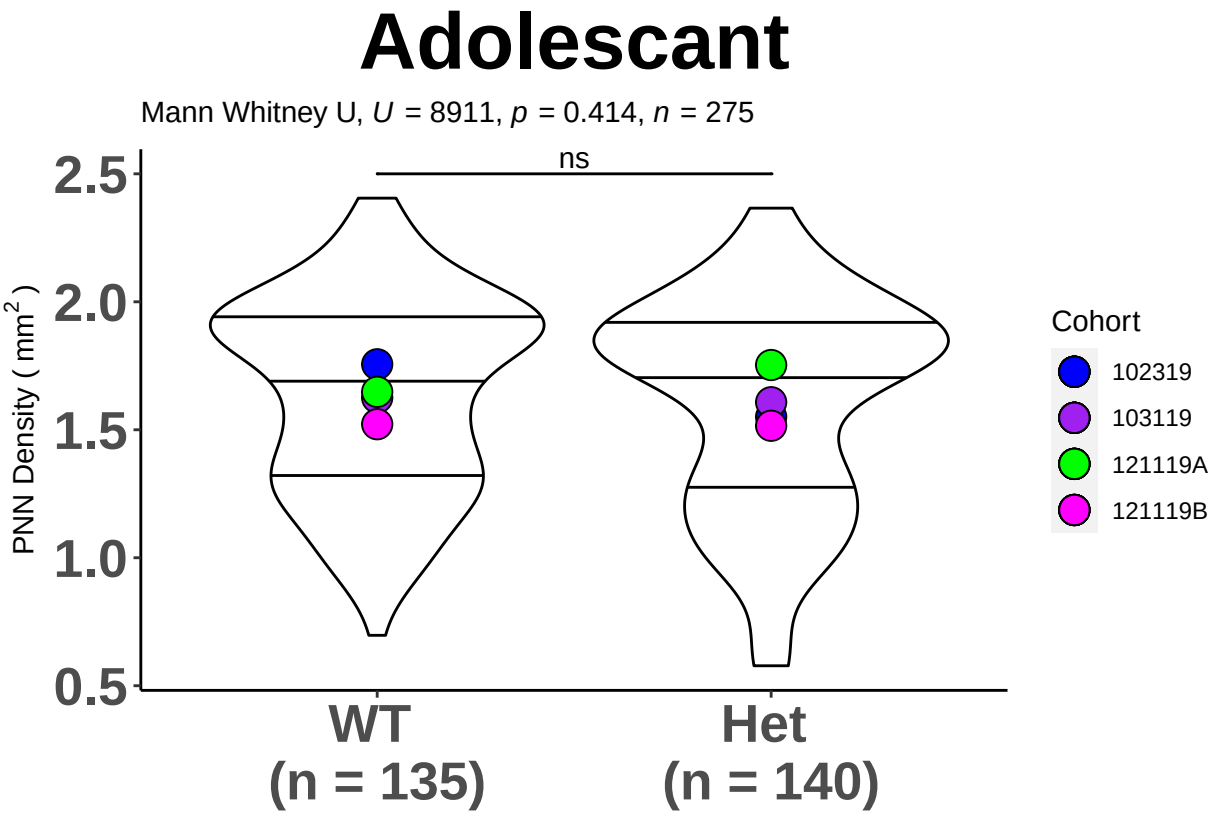
Figure 9c comparison. We are comparing the data we have to that in the pre-print to ensure consistency

```
##     X Cohort Condition Hemisphere Subregion Map.ID  Area  Weight.1  Weight.2
## 1 147 102319        WT       Left      S1BF     29 1.250 0.6235009 0.3764991
## 2 148 102319        WT       Left      S1BF     32 1.335 0.3381178 0.6618822
## 3 149 102319        WT       Left      S1BF     32 1.273 0.4210402 0.5789598
## 4 150 102319        WT       Left      S1BF     32 1.514 0.2772633 0.7227367
## 5 151 102319        WT       Left      S1BF     33 1.621 0.3067939 0.6932061
## 6 152 102319        WT       Left      S1BF     35 1.895 0.3399320 0.6600680
##     Mean.1   Mean.2 Variance.1 Variance.2      CV.1      CV.2 Index
## 1 33.99810 72.14835  149.98444   567.6658 0.3602206 0.3302326   147
## 2 31.70975 75.73504   63.29630   570.1870 0.2508975 0.3152912   148
## 3 30.48670 69.95238   92.47583   671.0181 0.3154305 0.3703093   149
## 4 34.41096 74.94763   71.47141   501.3757 0.2456798 0.2987609   150
## 5 33.59143 74.97395   84.48543   475.2736 0.2736292 0.2907779   151
## 6 29.86194 75.10848   70.47529   666.4199 0.2811256 0.3437043   152
##   X..Delta..Mean X..Delta..Weight
## 1       38.15025       -0.2470018
## 2       44.02529        0.3237643
## 3       39.46568        0.1579195
## 4       40.53667        0.4454735
## 5       41.38252        0.3864121
## 6       45.24653        0.3201360
```

```
## # A tibble: 1 x 12
##   estimate .y.   group1 group2    n1    n2 statistic     p conf.low conf.high
```

```
## *    <dbl> <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl>    <dbl>     <dbl>
## 1  -0.0340 Area  Het    WT       140   135      8911 0.414   -0.109    0.0529
## # ... with 2 more variables: method <chr>, alternative <chr>
```

PNN Figure 9 Plot



Overall Conclusion

Checking the `Het` only samples reveals that they have an ICC of ~0.61 which is higher than the overall of ~0.50. Looking at just them alone indicates we would need a little over 9 total cohorts. In essence this means they would need an `n` of 4 more cohorts (for a total of 10) as the value is slightly over 9 but better to have more than not enough. The `WT` only samples also have a moderate ICC value of ~0.54. In turn, they would need a little over 10 total cohorts (5 more) to account for the data dependence. This is not a big concern though because the statistical analysis remains unchanged for the lme PV samples. We will have to decide if we recommend an overall number where both `Het` and `WT` samples are inter-mixed or to offer a recommendation based on `Het` and `WT` alone.