Statistical Significance of DMLs by genomic regions

Cemile Balkas

2025-07-03

1 Introduction

This report presents the results of Fisher's exact tests on differentially methylated loci (DMLs) across genomic features compared to all tested CpG sites.

Fisher's exact tests were used to compare hyper- and hypomethylated DMLs across features like transcription start sites, exons, and intergenic regions. The analysis includes both exclusive overlaps (each CpG counted once) and non-exclusive overlaps (allowing overlaps with multiple features).

The aim is to see if DMLs are randomly distributed or if they tend to cluster in certain regions.

1.1 Load Data

```
df_dml_feature_counts <- read.csv("dml_feature_overlap_output/overlap_counts_exclusive.csv")
total_cpg <- sum(df_dml_feature_counts$All_CpG_Count)
df_dml_feature_counts$Percent_of_Total_CpGs <- round(100 * df_dml_feature_counts$All_CpG_Count / total_
kable(df_dml_feature_counts, caption = "Exclusive counts per genomic feature")
```

Table 1: Exclusive counts per genomic feature

Feature	All_CpG_Count	HyperDMLs	HypoDMLs	Percent_of_Total_CpGs
tss	492557	13708	11190	1.81
downstream_1kb	268352	2846	12471	0.99
5utrs	341338	11626	9960	1.26
3utrs	653184	5657	22290	2.41
exons	1096878	17683	36828	4.04
introns	11534889	282730	605257	42.49
intergenic	12758595	593117	1064491	47.00

```
df_overlap_inclusive <- read.csv("dml_feature_overlap_output/overlap_counts.csv")
df_overlap_inclusive$Percent_of_Total_CpGs <- round(100 * df_overlap_inclusive$All_CpG_Count / total_cpg
kable(df_overlap_inclusive, caption = "Non-exclusive counts per genomic feature")
```

Table 2: Non-exclusive counts per genomic feature

Feature	All_CpG_Count	HyperDMLs	HypoDMLs	Percent_of_Total_CpGs
tss	492557	13708	11190	1.81
downstream_1kb	285537	3038	13062	1.05
5utrs	548166	17469	14011	2.02
3utrs	690312	5995	23960	2.54
exons	2395519	42456	76664	8.82

Feature	All_CpG_Count	HyperDMLs	HypoDMLs	Percent_of_Total_CpGs
introns	11635556	284544	608747	42.86
intergenic	12759559	593141	1064521	47.00

1.2 Summarise totals

```
# Get total background counts
total_cpg <- sum(df_dml_feature_counts$All_CpG_Count)
total_hyper <- sum(df_dml_feature_counts$HyperDMLs)
total_hypo <- sum(df_dml_feature_counts$HypoDMLs)

# Show totals
total_cpg
```

```
## [1] 27145793
```

```
total_hyper
```

```
## [1] 927367
```

```
total_hypo
```

```
## [1] 1762487
```

1.3 Fisher's Exact test with box analogy

To determine whether differentially methylated loci (DMLs) are enriched in specific genomic features (eg TSS, exons), the distribution of DMLs is compared to all tested CpG sites using Fisher's Exact Test.

This can be explained by this analogy:

- Imagine a box containing 1,000 colored balls, each representing a tested CpG site.
  - 200 balls are red, representing CpGs in a particular genomic feature (eg TSS).
  - 800 balls are white representing CpGs not in that feature.
- If a total of 300 balls are randomly drawn from the box to represent DMLs.
  - Among these there are 90 red balls — ie 90 DMLs fall in TSS.
  - The question: Is observing 90 red balls (DMLs in TSS) more than expected by chance?

If DMLs were randomly distributed, the expected number of red DMLs would be:

300 * (200 / 1000) = 60

90 > 60 but is this statistically significant?

Fisher's Exact Test calculates the probability of seeing 90 or more red balls in a random draw of 300, given the background proportions using the hypergeometric distribution.

For Fisher's test to be valid:

- CpGs must fall into mutually exclusive groups: each site is either a DML or not, and either in the feature or not.
- All CpGs must come from the same tested background set, not the whole genome.

In this analysis, Fisher's test is applied first to exclusive overlaps, where each CpG is assigned to only one feature, and then to hyper dmls in non-exclusive overlaps.

```
box_analogy_CpGs <- data.frame(
  CpG_in_tss = c(90, 110),
  CpG_not_in_tss = c(210, 590),
```

```
  row.names = c("DML", "Non-DML")
)

box_analogy_CpGs
```

```
##         CpG_in_tss CpG_not_in_tss
## DML             90            210
## Non-DML        110            590
```

```
fisher.test(box_analogy_CpGs)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  box_analogy_CpGs
## p-value = 4.764e-07
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.645304 3.202829
## sample estimates:
## odds ratio
##   2.296597
```

1.4 Fisher's test example: Hyper-DMLs in TSS

Total CpGs tested (total balls in the box): 27,145,793 CpGs in TSS (red balls): 492,557

Hyper-DMLs (balls drawn from the box): 927,367 Hyper-DMLs in TSS (red balls drawn): 13,708

927367 * (492557 / 27145793) = 16,827

But only 13,708 observed : fewer than expected... statistically significant?

```
# Totals
total_cpgs <- 27145793
total_hyper_dmls <- 927367
tss_cpgs <- 492557
tss_hyper_dmls <- 13708

# Calculate values
hyperdml_in_tss <- tss_hyper_dmls
hyperdml_not_in_tss <- total_hyper_dmls - hyperdml_in_tss

non_dml_total <- total_cpgs - total_hyper_dmls
non_dml_in_tss <- tss_cpgs - hyperdml_in_tss
non_dml_not_in_tss <- non_dml_total - non_dml_in_tss

# Create table
hyperdml_CpGs <- data.frame(
  CpG_in_tss = c(hyperdml_in_tss, non_dml_in_tss),
  CpG_not_in_tss = c(hyperdml_not_in_tss, non_dml_not_in_tss),
  row.names = c("HyperDML", "Non-DML")
)

hyperdml_CpGs
```

```
##          CpG_in_tss CpG_not_in_tss
```

```
## HyperDML      13708         913659
## Non-DML      478849       25739577
```

```
fisher.test(hyperdml_CpGs)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  hyperdml_CpGs
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.7927592 0.8204087
## sample estimates:
## odds ratio
##  0.8064758
```

1.5 Run Fisher's Test for Hyper DMLs

Are hypermethylated DMLs more common in this feature than we would expect by chance?

We test if the proportion of DMLs in this feature is significantly different from what you'd expect based on the overall CpG background...

```
options(scipen = 0) # turn off scientific notation

# Run Fisher's test for all features for HyperDMLs
fisher_results_hyper <- data.frame()

for (i in 1:nrow(df_dml_feature_counts)) {
  # table logic
  #               In_feature    Not_in_feature      Total
  # DML               a               b           total_hyper_dmls
  # Not_DML           c               d                .
  # Total     total_cpg_in_feature    .            total_cpg

  total_cpg <- sum(df_dml_feature_counts$All_CpG_Count)
  total_cpg_in_feature <- df_dml_feature_counts$All_CpG_Count[i]
  total_hyper_dmls <- sum(df_dml_feature_counts$HyperDMLs)
  total_hyper_dmls_in_feature <- df_dml_feature_counts$HyperDMLs[i]

  a <- total_hyper_dmls_in_feature
  b <- total_hyper_dmls - a
  c <- total_cpg_in_feature - a
  d <- (total_cpg - total_cpg_in_feature) - b

  contingency <- data.frame(
    In_Feature = c(a, c),
    Not_In_Feature = c(b, d),
    row.names = c("DML", "Not_DML")
  )

  test <- fisher.test(contingency)

  fisher_pval <- test$p.value
  odds_ratio <- as.numeric(test$estimate)
```

```
  fisher_results_hyper <- rbind(fisher_results_hyper, data.frame(
    Feature = df_dml_feature_counts$Feature[i],
    p_value = fisher_pval,
    OddsRatio = odds_ratio
  ))
}

kable(fisher_results_hyper, caption = "Fisher's test results for hypermethylated DMLs (exclusive feature
```

Table 3: Fisher's test results for hypermethylated DMLs (exclusive features)

Feature	p_value	OddsRatio
tss	0.0000000	0.8064758
downstream_1kb	0.0000000	0.3009054
5utrs	0.7435551	0.9968582
3utrs	0.0000000	0.2423706
exons	0.0000000	0.4528115
introns	0.0000000	0.5833581
intergenic	0.0000000	2.0497974

1.6 Run Fisher's Test for Hypo DMLs

```
fisher_results_hypo <- data.frame()

for (i in 1:nrow(df_dml_feature_counts)) {
  # table logic
  #               In_feature    Not_in_feature      Total
  # DML               a               b           total_hypo_dmls
  # Not_DML           c               d                .
  # Total     total_cpg_in_feature    .            total_cpg

  total_cpg <- sum(df_dml_feature_counts$All_CpG_Count)
  total_cpg_in_feature <- df_dml_feature_counts$All_CpG_Count[i]
  total_hypo_dmls <- sum(df_dml_feature_counts$HypoDMLs)
  total_hypo_dmls_in_feature <- df_dml_feature_counts$HypoDMLs[i]

  a <- total_hypo_dmls_in_feature
  b <- total_hypo_dmls - a
  c <- total_cpg_in_feature - a
  d <- (total_cpg - total_cpg_in_feature) - b

  contingency <- data.frame(
    In_Feature = c(a, c),
    Not_In_Feature = c(b, d),
    row.names = c("DML", "Not_DML")
  )

  test <- fisher.test(contingency)

  fisher_pval <- test$p.value
```

```
  odds_ratio <- as.numeric(test$estimate)

  fisher_results_hypo <- rbind(fisher_results_hypo, data.frame(
    Feature = df_dml_feature_counts$Feature[i],
    p_value = fisher_pval,
    OddsRatio = odds_ratio
  ))
}

kable(fisher_results_hypo, caption = "Fisher's test results for hypomethylated DMLs (exclusive features)
```

Table 4: Fisher's test results for hypomethylated DMLs (exclusive features)

Feature	p_value	OddsRatio
tss	0	0.3305424
downstream_1kb	0	0.6997752
5utrs	0	0.4296473
3utrs	0	0.5025644
exons	0	0.4896871
introns	0	0.6916617
intergenic	0	1.7852292

```
# Rename columns for clarity before merging
names(fisher_results_hyper) <- c("Feature", "Hyper_p", "Hyper_Odds")
names(fisher_results_hypo) <- c("Feature", "Hypo_p", "Hypo_Odds")

# Merge on Feature
fisher_combined <- merge(fisher_results_hyper, fisher_results_hypo, by = "Feature")

# View combined results
kable(fisher_combined, caption = "Combined Fisher's test results for hyper- and hypomethylated DMLs (ex
```

Table 5: Combined Fisher's test results for hyper- and hypomethylated DMLs (exclusive features)

Feature	Hyper_p	Hyper_Odds	Hypo_p	Hypo_Odds
3utrs	0.0000000	0.2423706	0	0.5025644
5utrs	0.7435551	0.9968582	0	0.4296473
downstream_1kb	0.0000000	0.3009054	0	0.6997752
exons	0.0000000	0.4528115	0	0.4896871
intergenic	0.0000000	2.0497974	0	1.7852292
introns	0.0000000	0.5833581	0	0.6916617
tss	0.0000000	0.8064758	0	0.3305424

The Fisher's exact tests compare the proportion of hyper- and hypomethylated DMLs within each genomic feature to the background of all tested CpG sites.

1.6.1 Intergenic regions

Show strong enrichment for both hyper- and hypomethylated DMLs: - Odds Ratios: Hyper = 2.05, Hypo = 1.79 - This suggests that DMLs are overrepresented in intergenic regions, relative to the CpG background.

1.6.2 Exons, Introns and 3'UTRS

Show consistent depletion of DMLs: - Exons: Hyper = 0.45, Hypo = 0.49 - Introns: Hyper = 0.58, Hypo = 0.69 - 3' UTRs: Hyper = 0.24, Hypo = 0.50 These odds ratios suggest that DMLs are less likely to occur within gene bodies and 3' regulatory elements than expected by chance.

1.6.3 Transcription Start Sites

Show mild depletion for hyper-DMLs (OR = 0.81) and a stronger depletion for hypo-DMLs (OR = 0.33), both statistically significant.

This indicates that DMLs are less likely to be found in TSS regions, especially hypomethylated ones, suggesting that TSSs are relatively protected from methylation changes.

1.6.4 5' UTRs

Show mixed results:

- Hyper-DMLs: OR = 1.00, p = 0.74
  - no significant difference from background
- Hypo-DMLs: OR = 0.43
  - significant depletion

This indicates that hypermethylation in 5' UTRs occurs at background levels, while hypomethylation is significantly underrepresented.

1.7 Tests for overlaps wihtout prececence

```
options(scipen = 0) # turn off scientific notation

# Run Fisher's test for all features for HyperDMLs
fisher_results_inclusive_hyper <- data.frame()

for (i in 1:nrow(df_overlap_inclusive)) {
  # table logic
  #               In_feature    Not_in_feature      Total
  # DML               a               b           total_hyper
  # Not_DML           c               d                .
  # Total       total_in_feature        .            total_cpg

  a <- df_overlap_inclusive$HyperDMLs[i]
  total_in_feature <- df_overlap_inclusive$All_CpG_Count[i]
  c <- total_in_feature - a

  total_hyper <- sum(df_overlap_inclusive$HyperDMLs)
  total_cpg <- sum(df_overlap_inclusive$All_CpG_Count)

  b <- total_hyper - a
  d <- (total_cpg - total_in_feature) - b

  contingency <- matrix(c(a, b, c, d),
    nrow = 2,
    dimnames = list(
      Region = c("In_Feature", "Not_In_Feature"),
      Status = c("DML", "Not_DML")
    )
```

```
  )

  test <- fisher.test(contingency)

  fisher_pval <- test$p.value
  odds_ratio <- as.numeric(test$estimate)

  fisher_results_inclusive_hyper <- rbind(fisher_results_inclusive_hyper, data.frame(
    Feature = df_dml_feature_counts$Feature[i],
    p_value = fisher_pval,
    OddsRatio = odds_ratio
  ))
}

kable(fisher_results_inclusive_hyper, caption = "Fisher's test results for hypermethylated DMLs (non-ex
```

Table 6: Fisher's test results for hypermethylated DMLs (non-exclusive features)

Feature	p_value	OddsRatio
tss	0	0.8276208
downstream_1kb	0	0.3096455
5utrs	0	0.9536381
3utrs	0	0.2493399
exons	0	0.5011264
introns	0	0.6118821
intergenic	0	2.0818213

Fisher's exact tests were performed using both exclusive and inclusive genomic feature overlaps. Odds ratios were very similar across both approaches, meaning a consistent pattern of DML enrichment or depletion.

While p-values differed slightly most likely due to overlapping feature counts but the overall conclusions remain the same

1.8 Conclusions

To assess whether differentially methylated loci (DMLs) preferentially occur in specific genomic regions, I performed Fisher's exact tests comparing the distribution of hyper- and hypomethylated DMLs across annotated features relative to the background of all tested CpG sites.

The results showed extremely small p-values (0) for most genomic features, indicating highly significant differences between the observed distribution of DMLs and the background. Based on the odds ratios, DMLs were strongly enriched in intergenic regions, while significantly depleted in the majority of the rest of the regions.

These findings suggest that differential methylation is not randomly distributed throughout the genome, but tends to occur more often in certain genomic features.

Differentially methylated regions (DMRs) were identified using the same statistical thresholds and parameters as the DML analysis, i.e. the same minimum methylation difference, smoothing, and significance cut-offs.

Although permutation-based methods could be used in principle to assess the significance of DMRs, they are computationally intensive and were not applied here due to time constraints. However, since our Fisher's exact tests demonstrated that DMLs detected under the same parameters are highly statistically significant across multiple genomic features, we can infer that the DMRs identified are also likely to be satisfactory.

1.9 Testing how Fisher's Test work...

```
# my_test_df <- data.frame(
#   Background = c(492557, 27145793 - 492557), HypoDML = c(11190, 1762487 - 11190),
#   row.names = c("TSS", "NOT-TSS")
# )
#
# my_test_fisher <- fisher.test(my_test_df)
# my_test_fisher
#
#
# my_test_fisher$p.value
# my_test_fisher$estimate
#
# my_test_df <- data.frame(
#   Background = c(12758595, 27145793 - 12758595), HypoDML = c(1064491, 1762487 - 1064491),
#   row.names = c("TSS", "NOT-TSS")
# )
#
# my_test_fisher <- fisher.test(my_test_df)
# my_test_fisher
#
#
# my_test_fisher$p.value
# my_test_fisher$estimate
#
#
# # box analogy
# box_analogy_total_cpg <- 1000
# red_cpg_in_tss <- 200
# white_cpg_not_in_tss <- 800
#
# # randomly draw 300 balls (DMLs) and count what you observed
# red_dml_in_tss <- 90
# white_dml_not_in_tss <- 210
#
# # now we ask the quesion is 90 out of 300 more than what you'd expect by chance?
# # given only 200 out of 1000 are red...
# # lets calculate expected # of TSS dmls = 300 * (200/1000) = 60
# # 90 > 60 but is this difference statistically significant?
#
# p_value <- phyper(90 - 1, 200, 800, 300, lower.tail = FALSE)
# print(p_value)
#
# box_analogy_CpGs <- data.frame(
#   CpG_in_tss = c(90, 110), CpG_not_in_tss = c(210, 590),
#   row.names = c("DML", "Non-DML")
# )
#
# fisher.test(box_analogy_CpGs)
# fisher.test(box_analogy_CpGs)$p.value
```