Methods for the detection of conservation of methylation (R21, year 2)

(or, how statisticians can make your life more complicated)
Let’s detect conservation!

1. The phenotype of a cell is determined by its epigenome.
2. Develop methods to rank genes/pathways/regions according to their degree of conservation of methylation status.
3. Thus, we will identify the genes/pathways most important for growth of an individual Colorectal cancer
4. (A small step towards) Personalized medicine!

Darryl Shibata
That sounds great, but…

1. What exactly do you mean by conservation? How do we measure it?
2. How does your technology work?
3. What kind of error rates do you get using your technology (Illumina Infinium MethylationEPIC BeadChip Kit)?
4. You need this by when?!?

“Costs are skyrocketing. I don’t know how to feel about that.”
What does conservation mean?

<table>
<thead>
<tr>
<th></th>
<th>DNA methylation level</th>
<th>FDRP</th>
<th>qFDRP</th>
<th>PDR</th>
<th>MHL</th>
<th>Epipolymorphism</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reads</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0/1</td>
<td>1</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.6</td>
<td>0.6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.733</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>0.555</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.5</td>
<td>0.117</td>
<td>0.083</td>
<td>0.358</td>
<td>1</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.83</td>
<td>0</td>
<td>0</td>
<td>not defined</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.65</td>
<td>0</td>
<td>0</td>
<td>not defined</td>
</tr>
</tbody>
</table>

Figure 2. Comparison of average DNA methylation level and WSH scores for different sequencing read configurations.

_Scherer, et al., Nucleic Acids Research, 2020, Vol. 48, No. 8_
What does conservation mean?

**Figure 2.** Comparison of average DNA methylation level and WSH scores for different sequencing read configurations.

*Scherer, et al., Nucleic Acids Research, 2020, Vol. 48, No. 8*
Pooled Data

Our preliminary data are pooled, rather than cell-level.

• Output = “beta values”.
• After much QC, we can think of the beta value for a given site as the “proportion of cells that are methylated at that site in that sample”.

Samples from left and right side of tumor.
Statistics to measure conservation

- Variance - by site or by region
- Manhattan distance
- Proportion of sites with extreme methylation proportions ([0,0.2] or [0.8,1])
Overall strategy

- e.g., for a gene-based analysis:
  - Calculate the observed statistic value for each gene
  - Rank genes according to those values
  - Pick-off the genes with highest(lowest) value, as the most conserved.
• Look for genes in which methylation is statistically significantly conserved across normal tissue.

• Belief: those genes are “important.”

• Validate: Look at expression of those conserved genes in the Expression Atlas database. Is it high?

• Colon
• Small intestine
• Endometrium

Proof of concept / QC
Variance of Statistic

• Suppose \( n \) sites in the region/gene.

• Proportion of sites with extreme methylation proportions ([0,0.2] or [0.8,1])
  • Binomial distribution: variance is \( O(1/n) \)

• Variance
  • Variance of variance:
    
    \[
    \frac{(n - 1)^2 \mu_4}{n^3} - \frac{(n - 1)(n - 3)\mu_2^2}{n^3} = O\left(\frac{1}{n}\right)
    \]
Bootstrapping:

- For a gene with m CpG sites:
  - Construct a large number of ‘null genes’ that also have m CpG sites.
  - Calculate statistic value for each null gene.
  - Rank the statistic value for the observed gene within the set of values for null genes to get a p-value.
- Do this for all genes of all lengths, and look at those with lowest resulting p-values.

Emil Hvitfeldt, MS¹; Chao Xia, BSc¹; Kimberly D. Siegmund, PhD¹; Danyl Shibata, MD²; and Paul Marjoram, PhD¹

JCO Clin Cancer Inform 4:100-107. © 2020 by American Society of Clinical Oncology

---

**README.md**

**MethCon5**

The goal of methcon5 is to identify and rank CpG DNA methylation conservation along the human genome. Specifically, it includes bootstrapping methods to provide ranking which should adjust for the differences in length as without it short regions tend to get higher conservation scores.

The following repository includes an analysis in which this package was used.

**Installation**

Please note that the name of the package is in all lowercase.

You can install the released version of methcon5 from CRAN with:

```r
install.packages("methcon5")
```

github.com/USCbiostats/MethCon5
More complex models?

More complex models? 

\[ P_X(x) = \frac{1}{Z} \exp \left\{ \sum_{n=1}^{N} a_n (2x_n - 1) \right\} \]

\[ + \sum_{n=2}^{N} c_n (2x_n - 1)(2x_{n-1} - 1) \]

(1)

\[ a_n = \alpha_k + \beta_k \rho_n, \]

\[ c_n = \frac{\gamma_k}{d_n}, \]

\( \alpha_k \): region-specific tendency to be methylated

\( \rho_n \): density of CpGs at \( n^{th} \) CpG

\( \beta_k \): region-specific dependence on density

\( d_n \): distance between \( n^{th} \) CpG and \((n-1)^{th}\) CpG

\( \gamma_k \): region-specific correlation parameter
Software: DNAMeda (Shiny App - Visualization)

Samples are selected by specifying plate, well and sample id. Data is specified globally in the application.

Plotting window: Dynamically changes with data selection and input.

github.com/USCbiostats/DNAMeda
HiLDA = “Hierarchical Latent Dirichlet Allocation”

- Somatic mutations
- \textit{pmsignature}
- Estimated proportions, $\hat{q}$
- e.g., Wilcoxon
- Are $\hat{q}$ different in two groups?

Uncertainty in proportions:
$q^0 \sim Dir(p_{01}, \ldots, p_{0K}, \phi_0)$
$q^1 \sim Dir(p_{11}, \ldots, p_{1K}, \phi_1)$
HiLDA: a statistical approach to investigate differences in mutational signatures

Zhi Yang¹, Priyatama Pandey¹, Darryl Shibata², David V. Conti¹, Paul Marjoram¹ and Kimberly D. Siegmund¹

¹ Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States of America
² Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States of America

HiLDA: "Hierarchical Latent Dirichlet Allocation"

Introduction

The R package HiLDA is developed under the Bayesian framework to allow statistically testing whether there is a change in the mutation burdens of mutation signatures between two groups. The mutation signature is defined based on the independent model proposed by Shiraishi's et al.

Paper


github.com/USCbiostats/HiLDA
Software: iMutSig
iMutSig - Shiny app.
Acknowledgements

USC: Emil Hvorfeldt, Kim Siegmund, Darryl Shibata, Zhi Yang

JH: Hari Easwared, Tom Pisanic

And many thanks to ITCR