Sequence, Bisulfite, Machine Learning and Mutation

Youqing Xiang

July 19, 2017

Introduction: Sequence

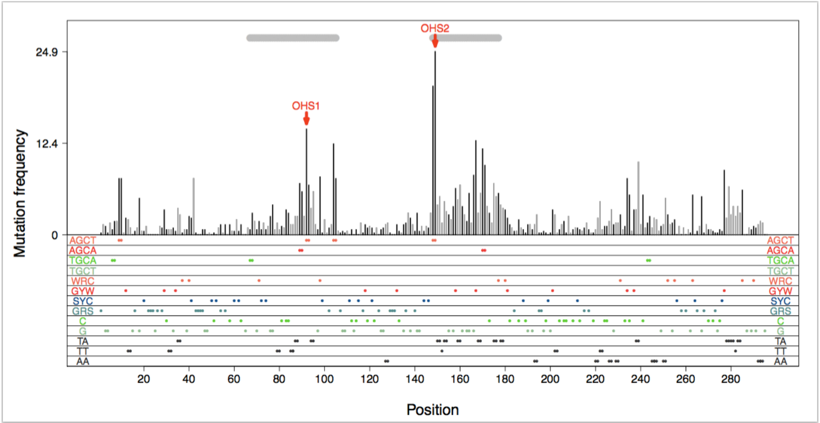
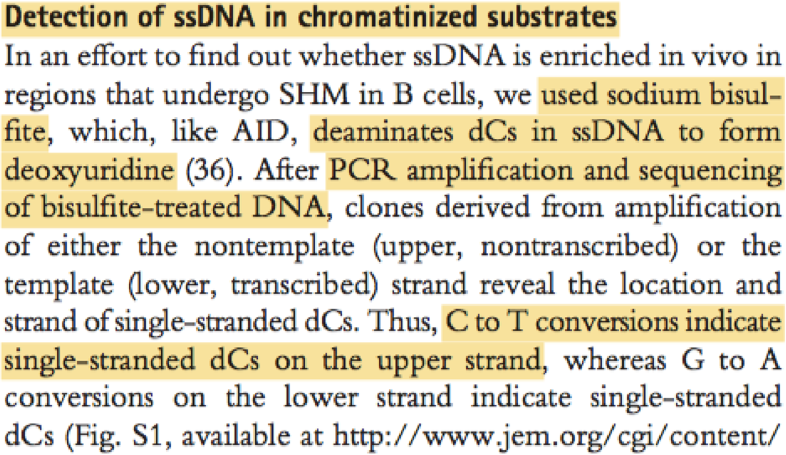


Figure 1:

From Wei, L etal PNAS 2015

Introduction: Bisulfite



Detection of ssDNA in chromatinized substrates

In an effort to find out whether ssDNA is enriched in vivo in regions that undergo SHM in B cells, we used sodium bisulfite, which, like AID, deaminates dCs in ssDNA to form deoxyuridine (36). After PCR amplification and sequencing of bisulfite-treated DNA, clones derived from amplification of either the nontemplate (upper, nontranscribed) or the template (lower, transcribed) strand reveal the location and strand of single-stranded dCs. Thus, C to T conversions indicate single-stranded dCs on the upper strand, whereas G to A conversions on the lower strand indicate single-stranded dCs (Fig. S1, available at http://www.jem.org/cgi/content/

Figure 2:

Introduction: Machine Learning

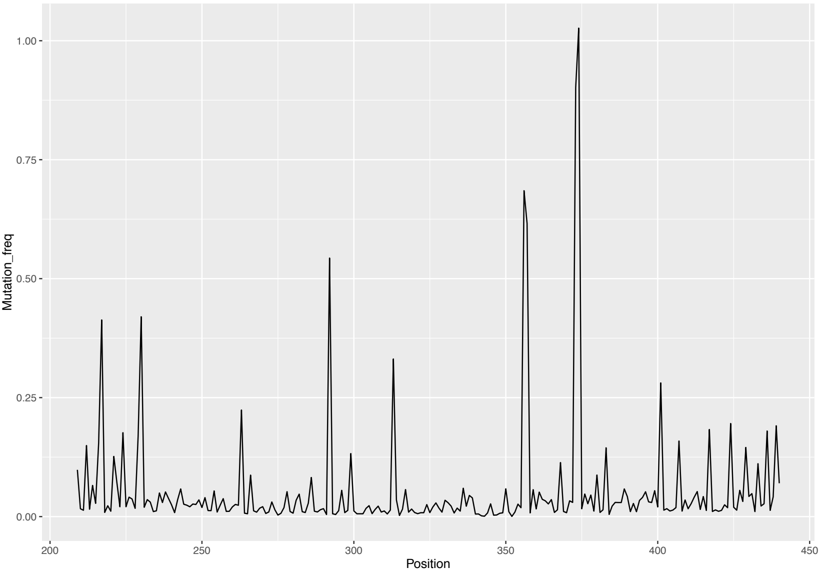
What is Machine Learning?

- Machine learning is a method of data analysis that automates analytical model building
- Using algorithms that iteratively learn from data, machine learing allows computers to find hidden insights without being explicitly programmed where to look.

What is it used for?

- Fraud detection
- New pricing models
- Financial Modeling
- Image recognition
- Text Sentiment Analysis
- . . . .

Introduction: Mutation from Ramos 4-34



Original Data Source and Tools

Data

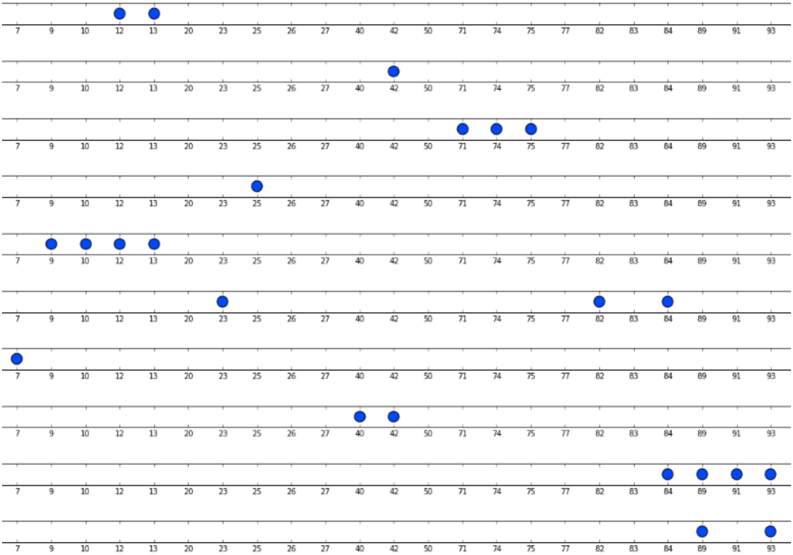
- Bisulfite.Alberto.fasta
- Ramos wild type 4-34

Tools

- Python and R
- Biopython, DNASTAR
- matplotlib and ggplot2
- randomForest

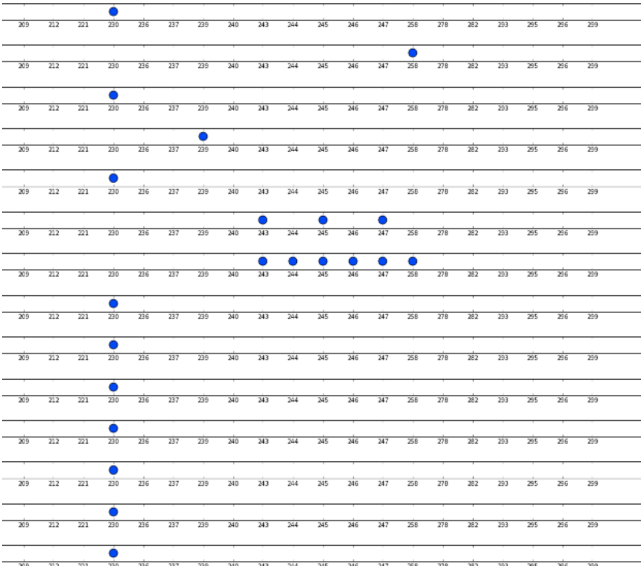
Bisulfite Data: Data Visulization

What is bisulfite accessible?



Bisulfite Data: Data Visulization

Bisulfite accessible site or region?



Bisulfite Data: Data Visulization

C to T Conversion

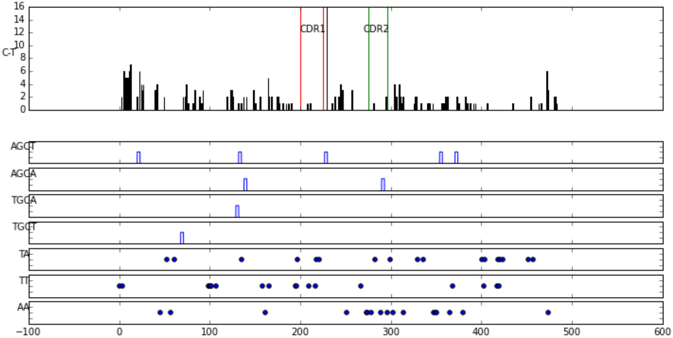


Figure 5:

Bisulfite Data: Data Visulization

G to A Conversion

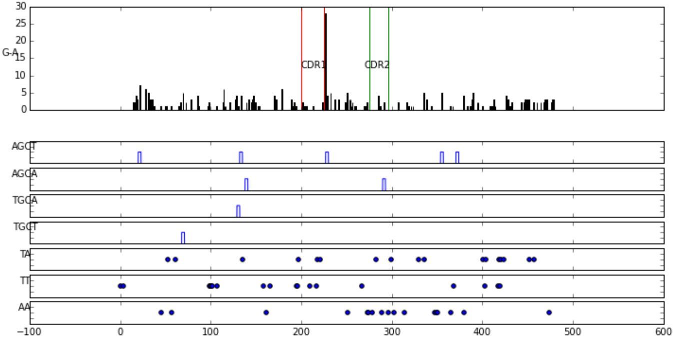


Figure 6:

Bisulfite Data: Future work

Study C to T (G to A) conversion sites

- Step 1: Dividing C/G positions into three groups (in hotspot, coldspot and neutral spot)
- Step 2: Calculate C to T conversion (G to A conversion) for each of three groups
- Step 3: Data visulization for the resuts from step 2
- Step 4: T test or ANOVA test for statistical significance

Study Bisulfite Acessible Region

- Collect more data
- Data Visulization and statistical analysis

Combined Data: Data Visulization

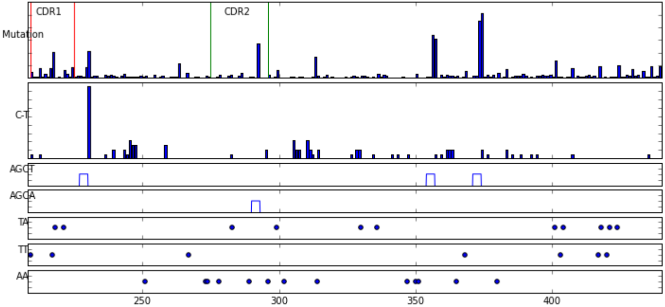
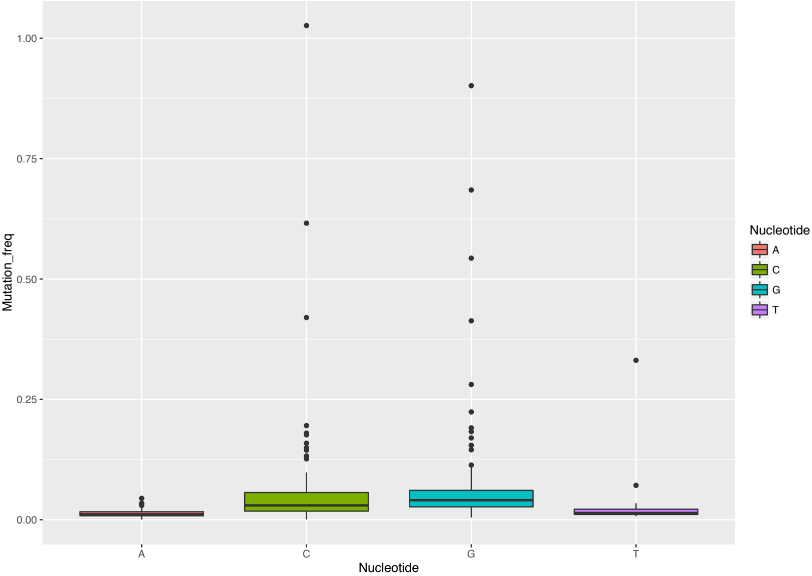


Figure 7:

Combined Data: Data Visulization



Combined Data: Machine Learning

Features

- Position
- Nucleotide type (G,C, A, T)
- Distance to AGCT (agct_d)
- Distance to AGCA (agca_d)
- Distance to TA (ta_d)
- Distance to TT (tt_d)
- Distance to AA (aa_d)
- C to T conversion rate (C_T)
- Distance to bisulfite accessible site (bisulf_d)

Labels

Mutation: High (1) and Low (0)

Combined Data: Machine Learning

Benchmark

Assuming all labels are 1.

Error rate: 50%

Combined Data: Machine Learning

randomForest

Training error rate: 21.6%

```
##            IncNodePurity
## Position        4.883839
## Nucleotide     17.175575
## agct_d          3.951250
## agca_d          6.152648
## ta_d            4.078553
## tt_d            5.991813
## aa_d            5.400212
## C_T             1.024082
## bisulf_d        3.784090
```

Combined Data: Future work

Collecting more data

- Sequence data
- Bisulfite data

Creating new features

Optimizing models and Evaluating models

Testing models with experiments