IG Modification

Yichen Han

2024-08-22

Flexible Baseline

We modified the source code of `deepG::integrated_gradients()` so that it now accepts any one-hot coded baseline array as an argument, thus allowing flexible comparison. It is currently only modified for the simplest case: no repeat, and input instance and baseline are not a list.

To showcase that a correctly chosen baseline can be crucial for model explanations, we load the same model as in RSDexport, and set two baselines:

1. natural baseline: the documented gene sequence of *rsd*.
2. abnormal baseline: a randomly functionally permuted copy, which should be predicted as "abnormal" by the model.

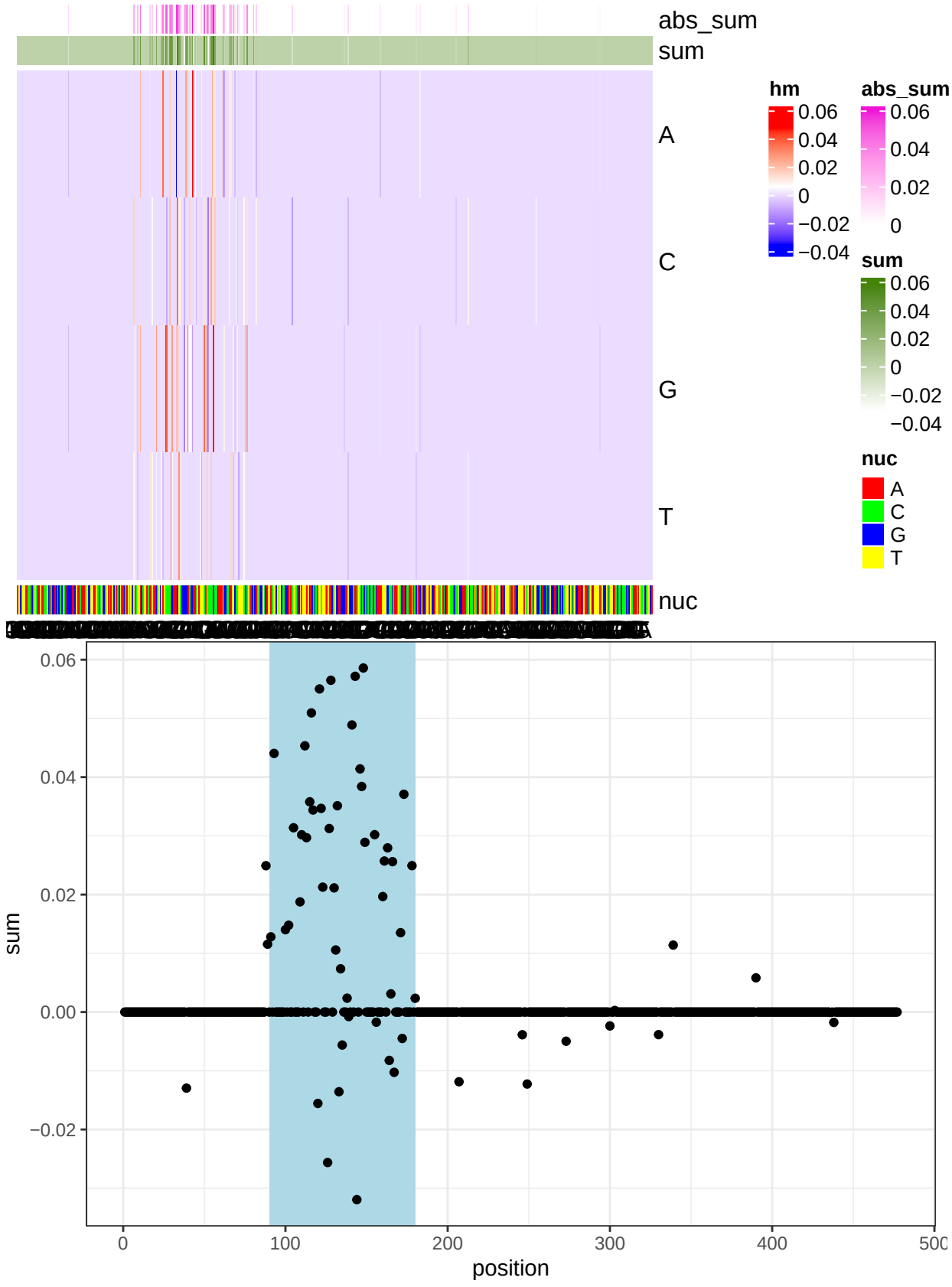
```
## Using checkpoint checkpoints/rsd-permutation_39/Ep.010-val_loss0.18-val_acc0.946.hdf5
```

After modification, the IG call is the following:

```
ig <- ig_modified(
  input_seq = onehot_instance,
  baseline_type = "modify",
  baseline_onehot = onehot_baseline,
  target_class_idx = 3,
  model = model,
  num_baseline_repeats = 1)
```

```
##              [,1]         [,2] [,3]
## [1,] 9.656643e-09 1.406469e-18    1
```

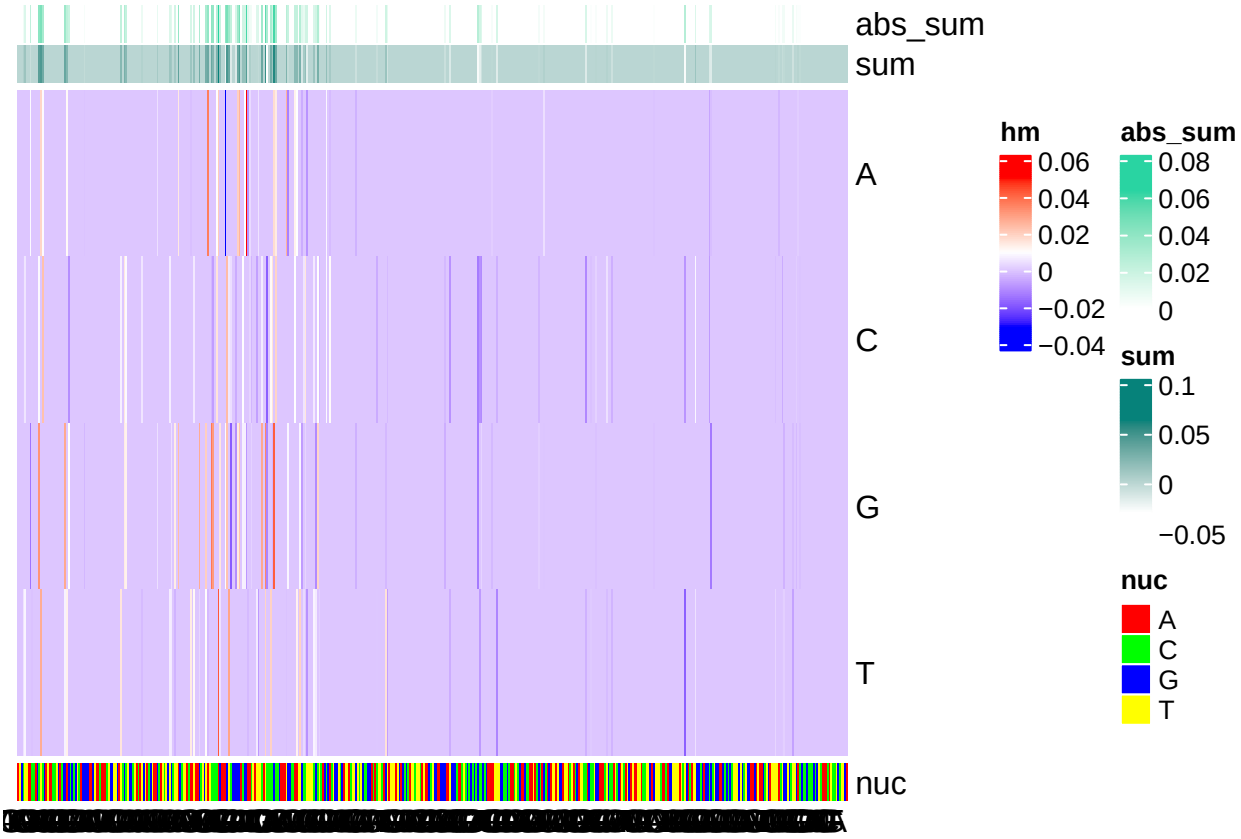
```
## [[1]]
```

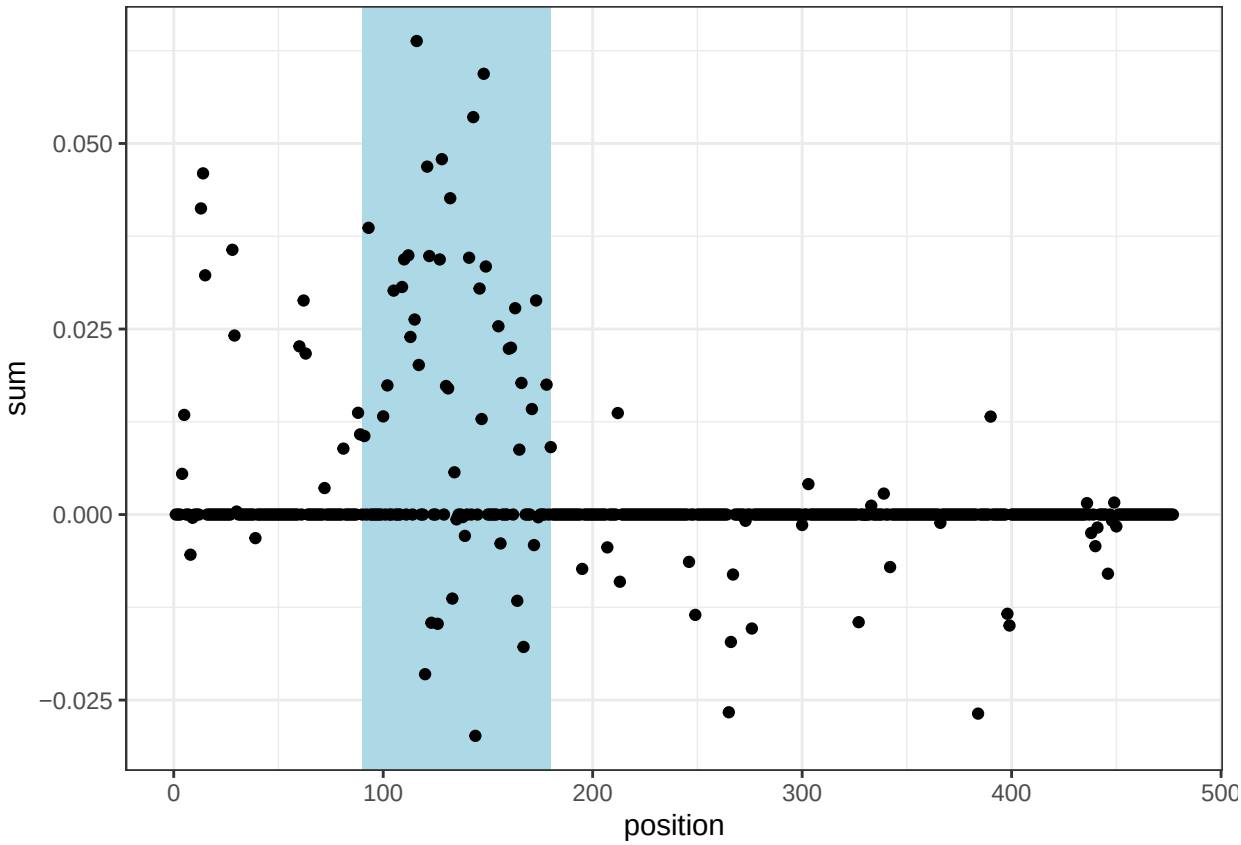


We see exactly what we expected – high importance assigned to the targeted area, while irrelevant positions

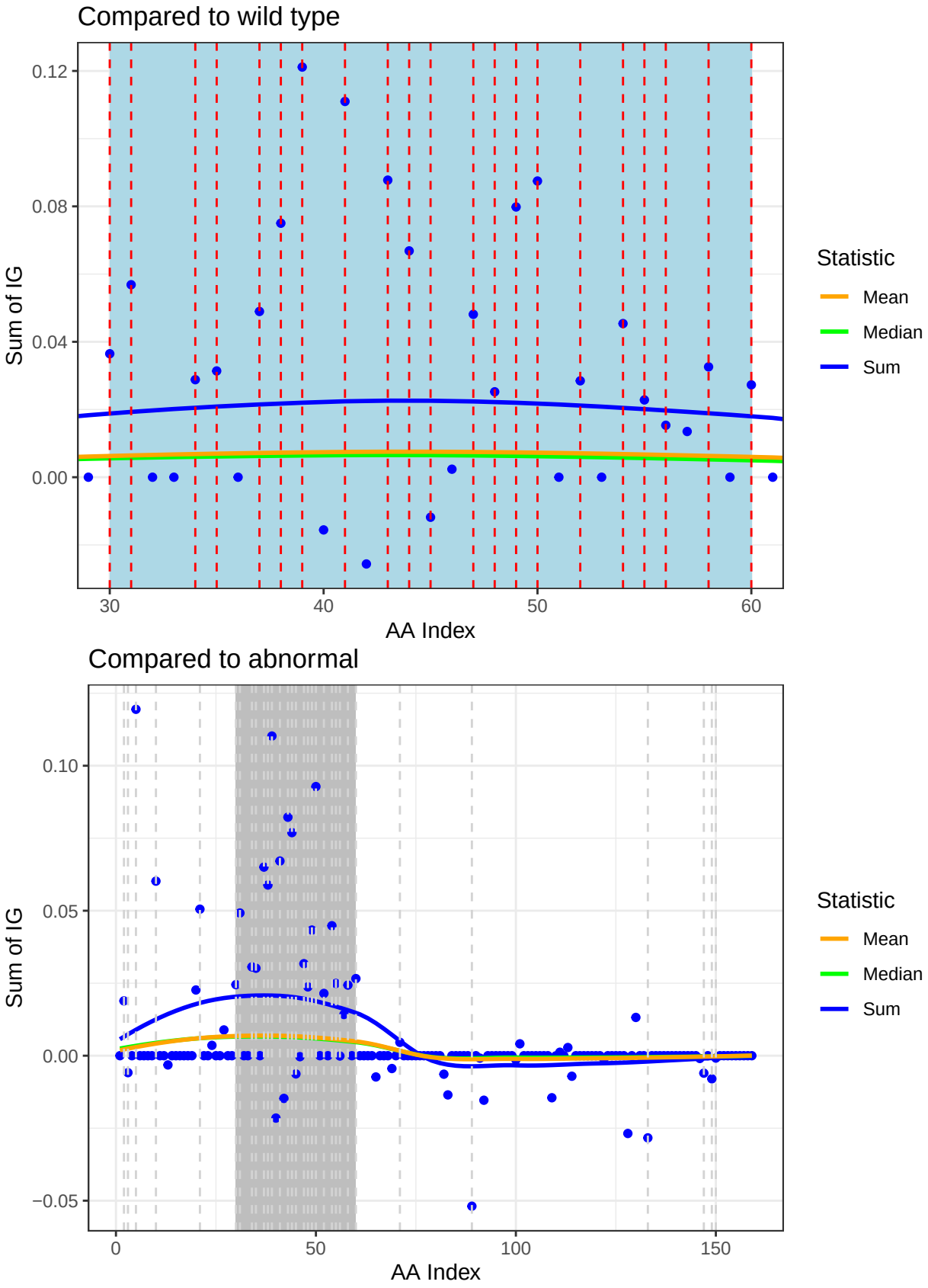
usually have gradients near 0.

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

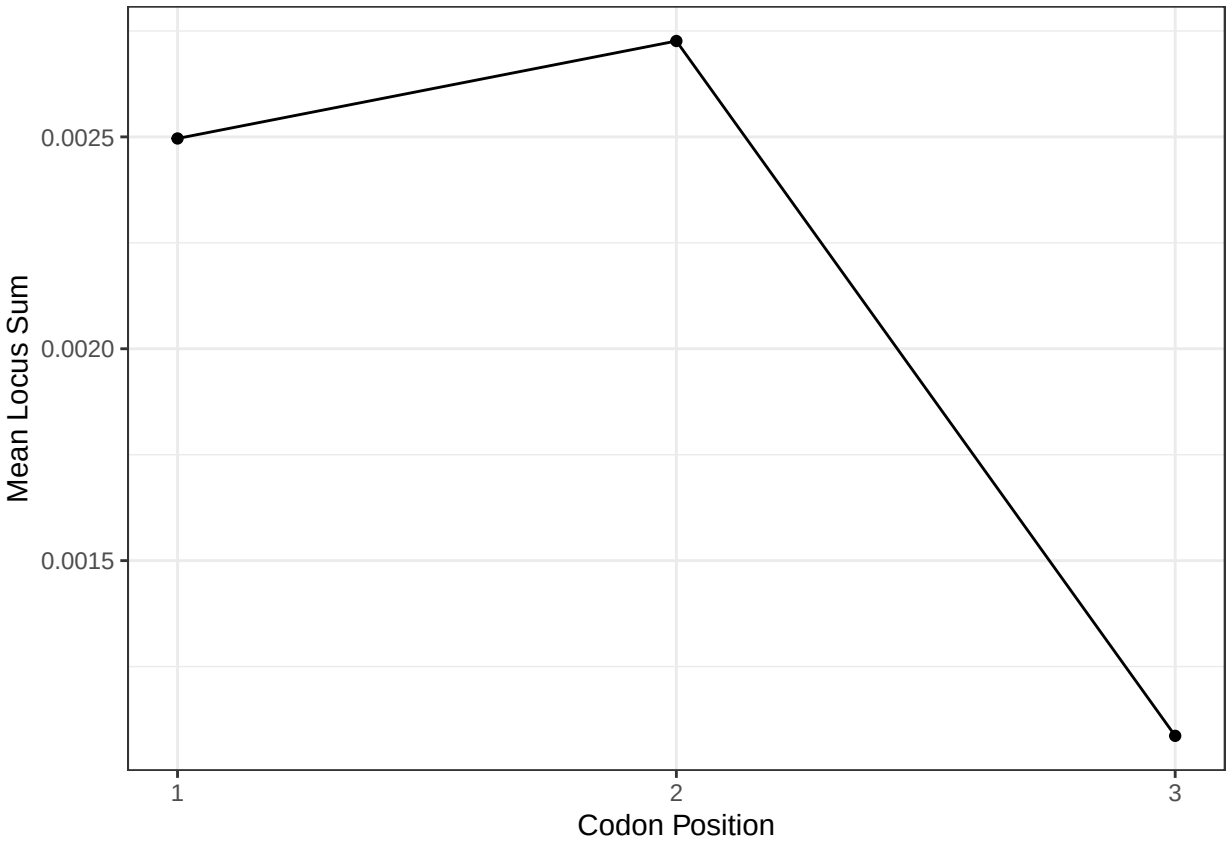


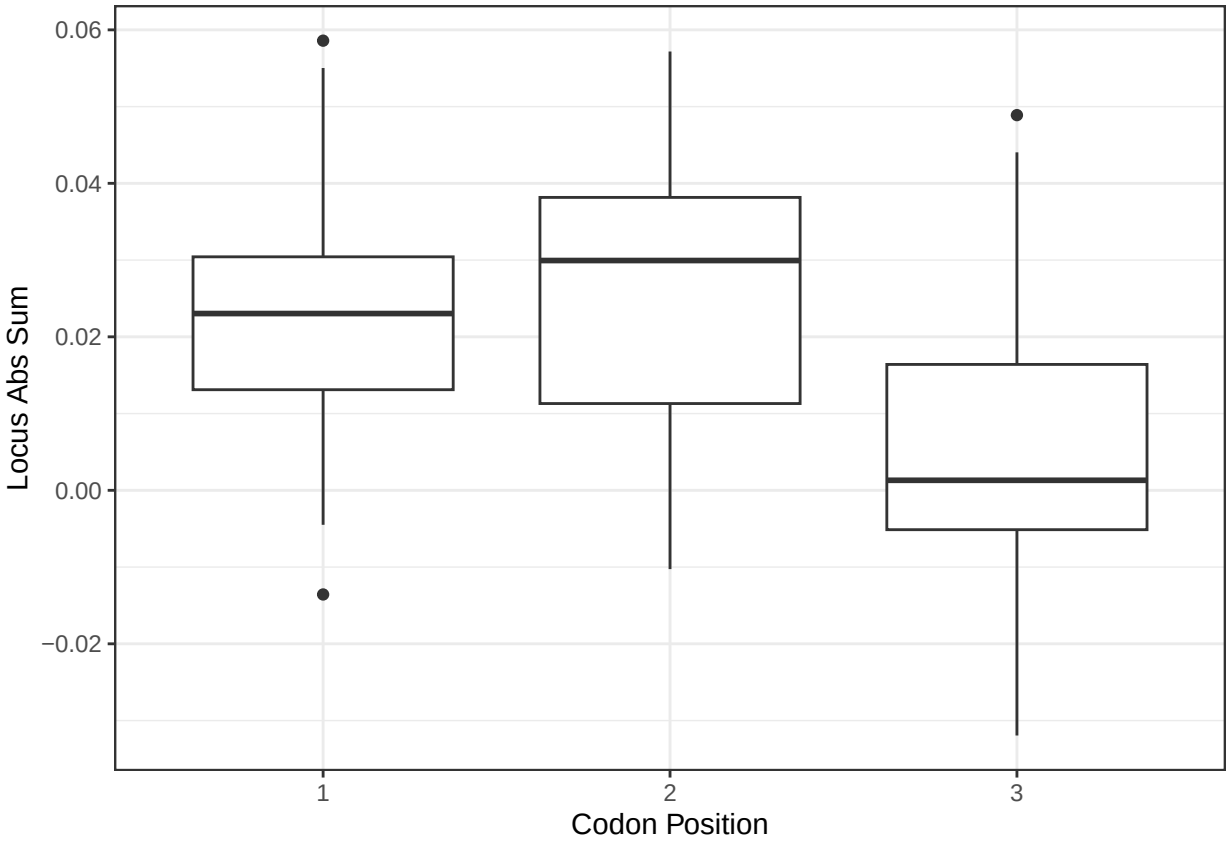


This is less clear when the baseline is the abnormal type, but we can see the methods clearly captures what matters and what does not.



We display it again after compressing the information. Vertical lines indicate substitution spots.


0.0025
0.0020
0.0015
Mean Locus Sum
1
2
3
Codon Position


0.06
0.04
0.02
0.00
−0.02
Locus Abs Sum
1
2
3
Codon Position

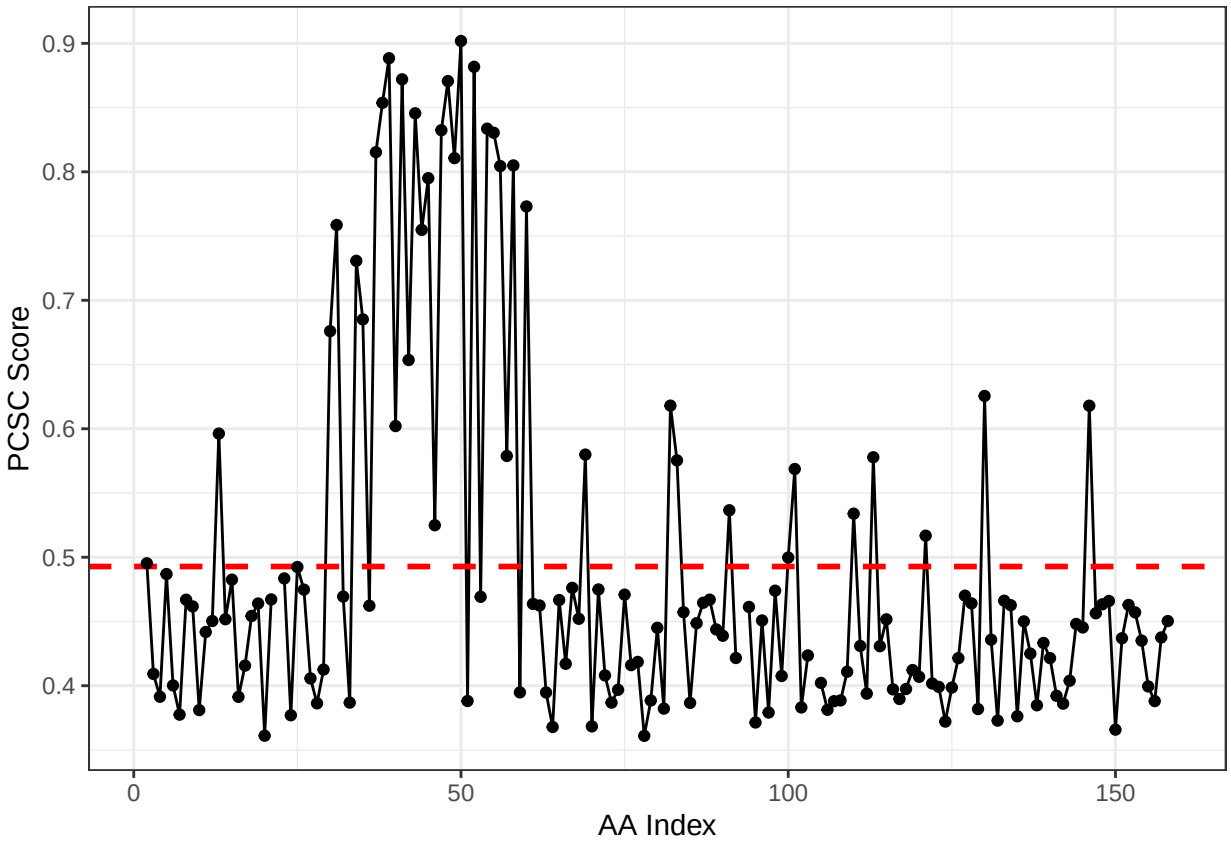
```
##
## Call:
## lm(formula = sum ~ factor(position), data = codon_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.038117 -0.011626 -0.001138  0.009794  0.042718
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.022050   0.004792   4.601 2.28e-05 ***
## factor(position)2   0.005042   0.006986   0.722   0.4733
## factor(position)3 -0.015880   0.006142  -2.585   0.0122 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02033 on 59 degrees of freedom
## Multiple R-squared:  0.1794, Adjusted R-squared:  0.1516
## F-statistic: 6.448 on 2 and 59 DF,  p-value: 0.002932
```

Wobbleness is again present, this time with relative confidence. Whether this effect is persistent across scenarios, is a topic worth further studying. But we need to recognize the high randomness behind it.

```
## Loaded result_df from existing CSV file.
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_line()`).
```

```
## Warning: Removed 5 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



We see that the model under correct explanation achieves much less consistency in treating synonymous codons than that of RSDexport report. The consistency is only about 0.5. Given the satisfactory size of data (15000), we speculate that the model can be improved in this regard.