

CAGI slide pack guidelines

I agree to my slides being used according to the following terms: Max Schubach

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.





CAGI 5

Single base-pair resolution variant effect readouts of disease associated promoter and enhancer elements

Max Schubach

BIH-JRG Computational Genome Biology

Chicago, IL, 2018-07-06

Contributors & funding



Jay Shendure
(UW)



Nadav Ahituv
(UCSF)



Ahituv & Shendure labs:

Beth Martin, Chenling Xiong & Fumitaka Inoue



Daniela Witten
(UW)

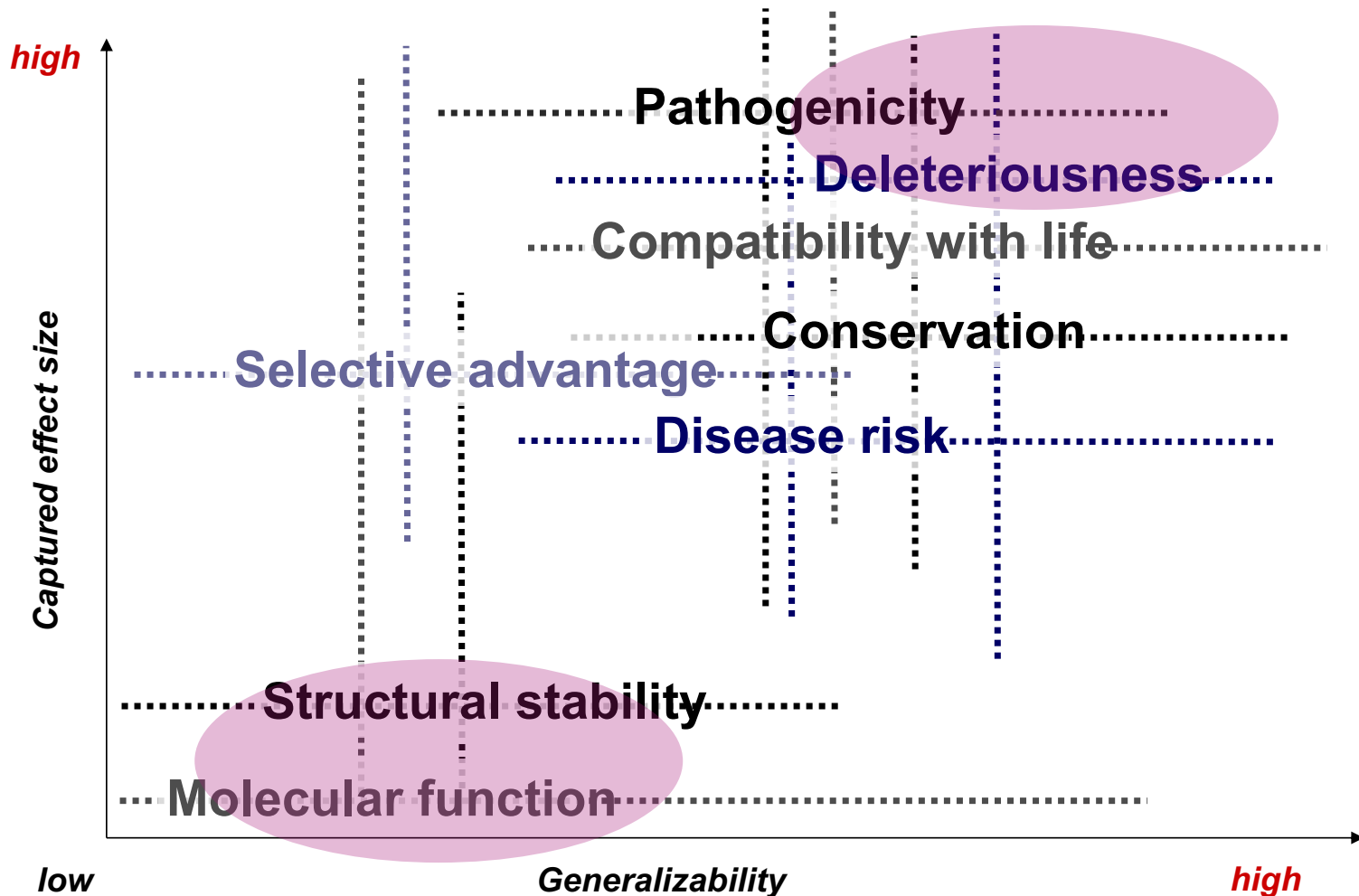


Martin Kircher
(BIH)

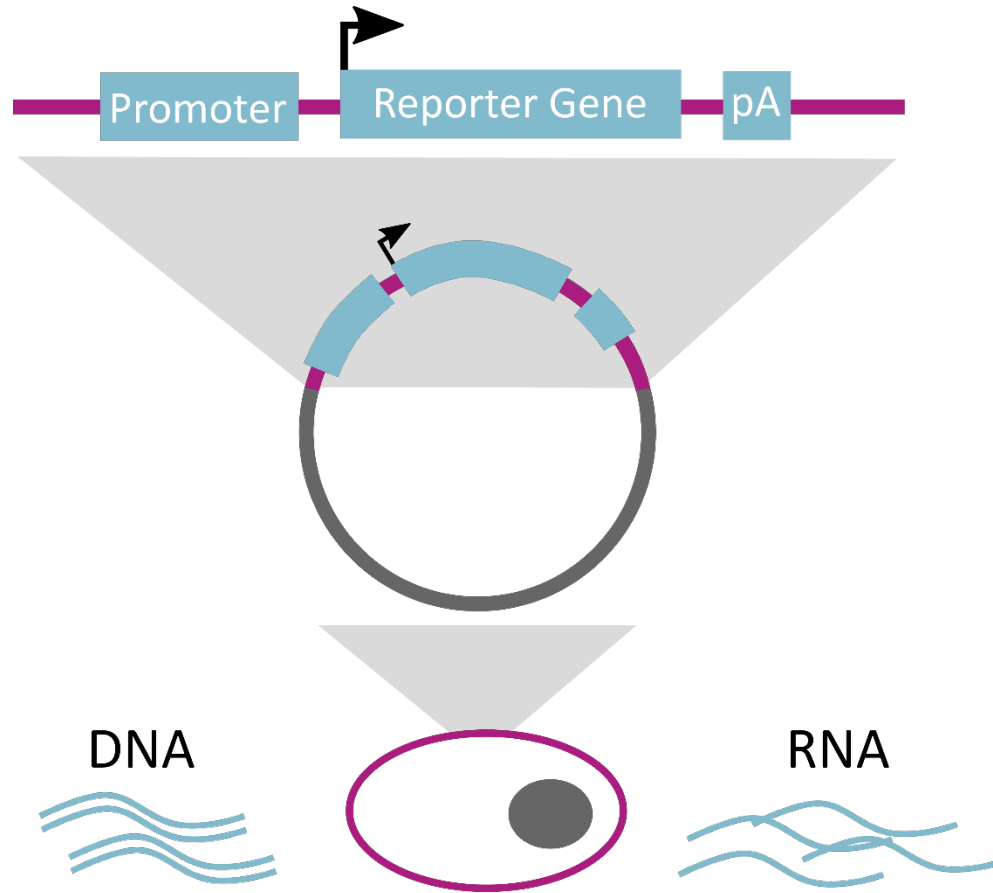


Funding:

More variants for a different effect size?

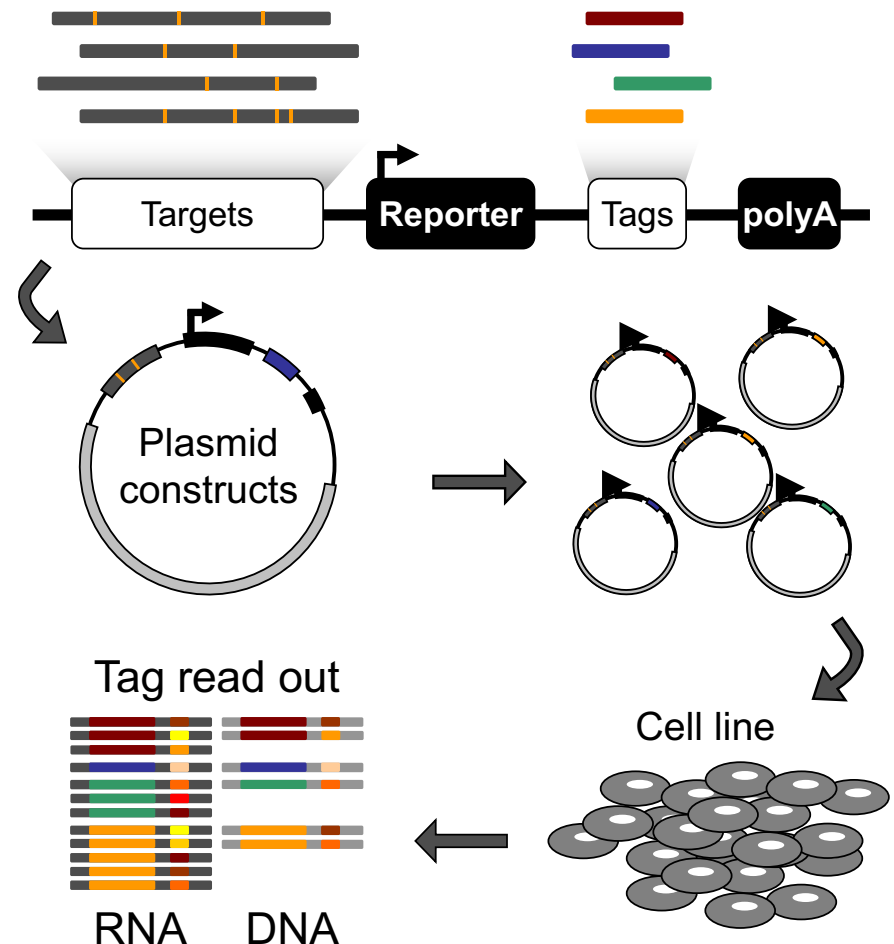


Reporter assays



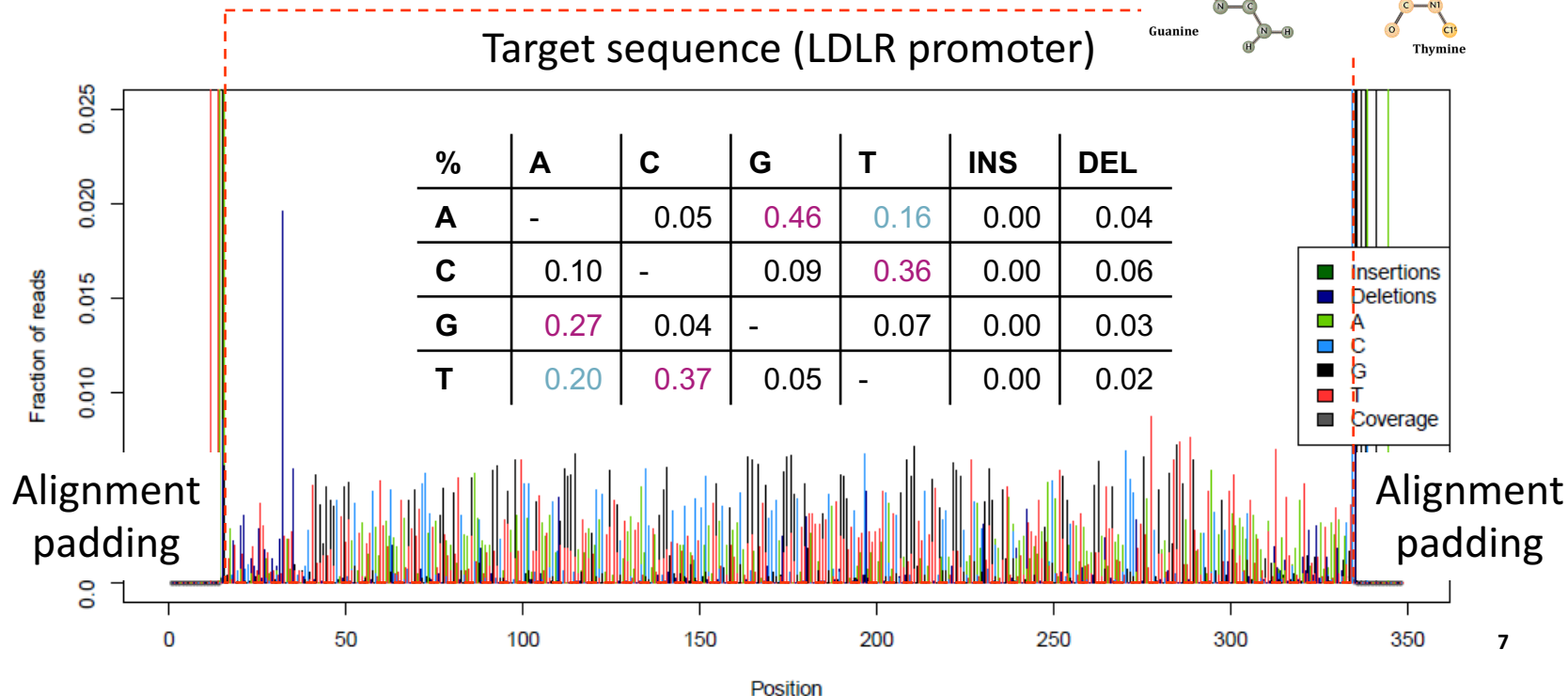
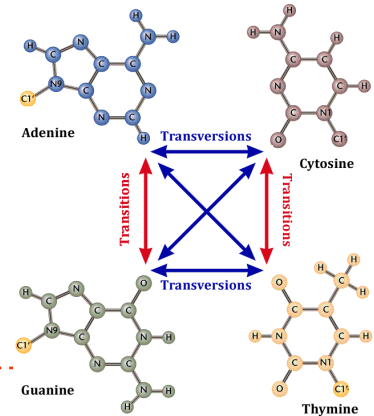
Saturation mutagenesis MPRA

- Error-prone PCR to generate sequence variants of regulatory region
- Integrate PCR products in plasmid library containing random tag sequences
- Learn association between tags and sequence variants (50k-2M)
- Express plasmids in cell line and collect RNA & DNA to sequence tags



Saturation mutagenesis

- Preference for transitions
- $T \leftrightarrow A$ is preferred transversion
- Obtained 100-160k tag associations easily cover all possible SNVs

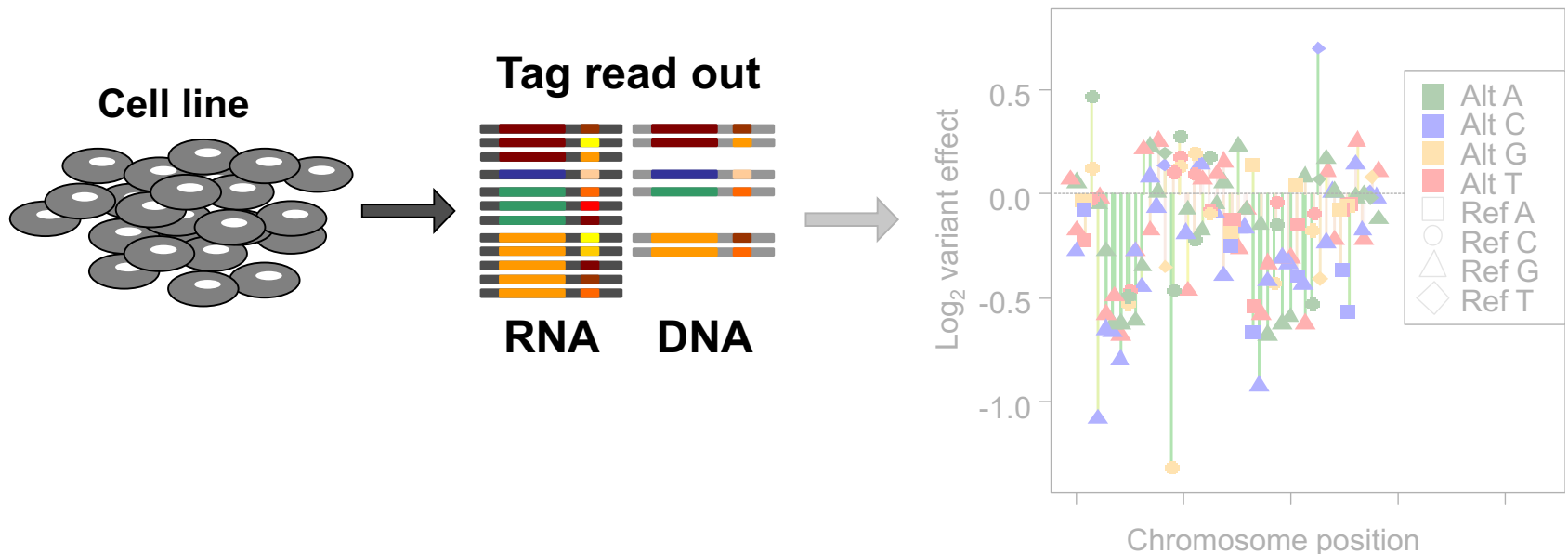


Selection of promoter and enhancers

- Selected 15 disease-associated promoter and enhancer elements, 187bp to 600bp in length, 28-73% GC content
- 9 promoters, e.g.
 - Telomerase reverse transcriptase (TERT, 259bp): Most frequent (known) non-coding mutations in cancer
 - Low Density Lipoprotein Receptor (LDLR, 318bp): Familial hypercholesterolemia
 - Coagulation Factor IX (F9, 303bp): Hemophilia B
- 5 enhancers, e.g.
 - Enhancer of Sortilin 1 gene (SORT1, 600bp): regulates low-density lipoprotein cholesterol (LDL-C)
 - Enhancer of Interferon Regulatory Factor 6 gene (IRF6, 601bp): Cleft lip

Measuring variant effects

- Count RNA/DNA molecules for each tag
- Generate a matrix V of variants (binary encoding of presence/absence) associated with the RNA and DNA tag counts
- Fit linear model of $\log_2(RNA) \sim \log_2(DNA) + V + intercept$, where the coefficients of V are the resulting individual variant effects



Measuring individual variant effects

- Count RNA/DNA molecules for each tag
- Generate a matrix V of variants (binary encoding of presence/absence) associated with the RNA and DNA tag counts
- Fit linear model of $\log_2(RNA) \sim \log_2(DNA) + V + intercept$, where the coefficients of V are the resulting individual variant effects

#Barcode	DNA	RNA	20_T.A	20_T.C	20_T.G	21_G.A	21_G.C	21_G.T	22_G.A
AAAAAAAAACATGTG	2	1	0	0	1	0	0	0	0
AAAAAAAAAGTGGCC	2	16	0	0	0	0	0	0	0
AAAAAAAAATTATAC	7	8	0	0	0	0	0	0	0
AAAAAAAAATTTCTCG	1	4	0	0	0	0	0	0	0
AAAAAACCTAGACA	3	3	0	1	0	0	0	0	0
AAAAAACGGAAATA	8	37	0	0	0	0	0	0	0
AAAAAAGATAGATG	9	32	0	0	0	0	0	0	0
AAAAAAGCGGGGGG	2	1	0	0	0	0	0	0	0
AAAAAAGTGTGGCA	1	5	0	0	0	0	0	0	0
AAAAAATCCTTGTG	2	6	0	0	0	0	0	0	0
AAAAAATGTCGTAG	2	14	0	0	0	0	0	0	0
AAAAAACAAATTTGC	3	11	0	0	0	0	0	0	10
AAAAACAAGCCGCT	2	3	0	0	0	0	0	0	0
AAAAACGCTGCTGCG	2	16	0	1	0	0	0	0	0

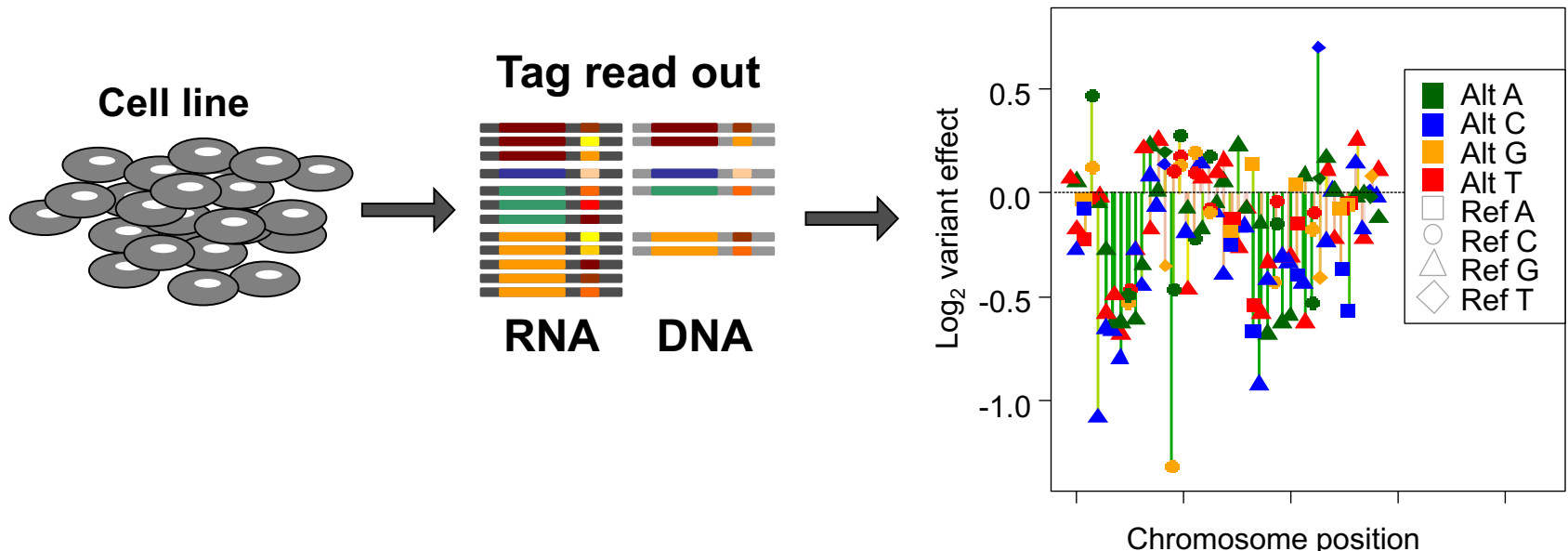
Measuring individual variant effects

- Count RNA/DNA molecules for each tag
- Generate a matrix V of variants (binary encoding of presence/absence) associated with the RNA and DNA tag counts
- Fit linear model of $\log_2(RNA) \sim \log_2(DNA) + V + intercept$, where the coefficients of V are the resulting individual variant effects

```
model <- glm(as.formula(sprintf("log2(RNA) ~ log2(DNA) +
%s",paste(colnames(V),collapse=" + "))))
...
# Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept) -2.387e-01  8.579e-03 -27.829  < 2e-16 ***
# X200_C.A      2.929e-01  5.617e-02   5.215  1.84e-07 ***
# X200_C.G      1.618e-01  1.443e-01   1.121  0.262280
# X200_C.T      2.280e-01  7.683e-02   2.968  0.002998 **
# X201_T.A      2.110e-01  8.502e-02   2.482  0.013065 *
# X201_T.C      2.026e-01  7.508e-02   2.698  0.006979 **
# X201_T.G      3.027e-01  1.392e-01   2.175  0.029597 *
# X202_G.A      1.297e-01  3.910e-02   3.317  0.000909 ***
# X202_G.C      1.899e-01  7.000e-02   2.713  0.006672 **
```

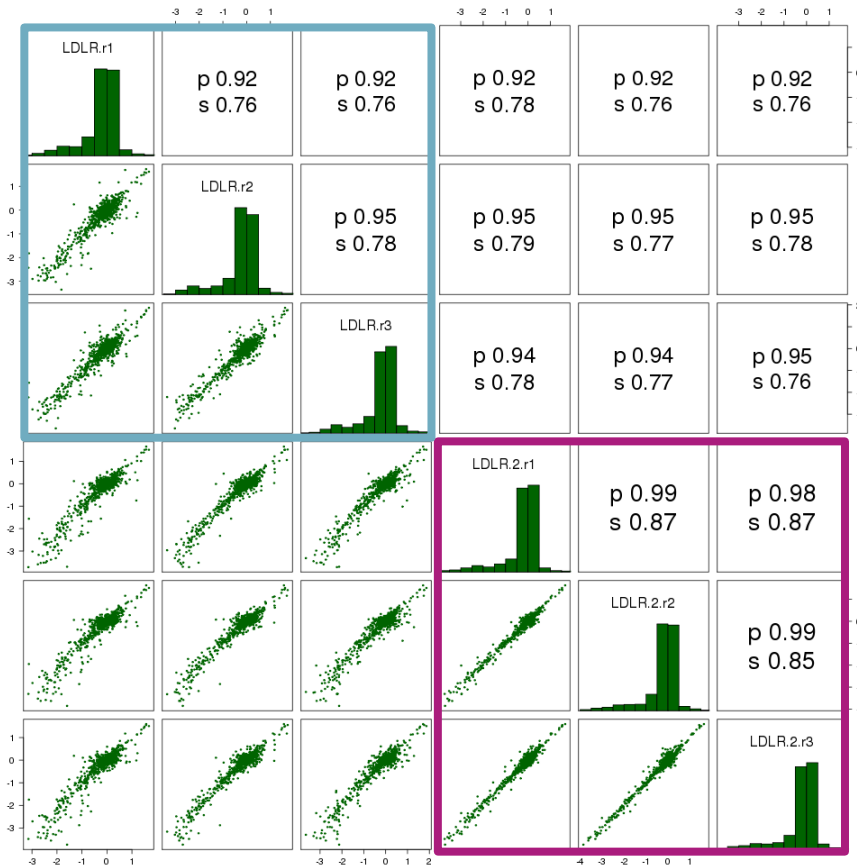
Measuring individual variant effects

- Count RNA/DNA molecules for each tag
- Generate a matrix V of variants (binary encoding of presence/absence) associated with the RNA and DNA tag counts
- Fit linear model of $\log_2(RNA) \sim \log_2(DNA) + V + intercept$, where the coefficients of V are the resulting individual variant effects



Reproducibility & experimental variation

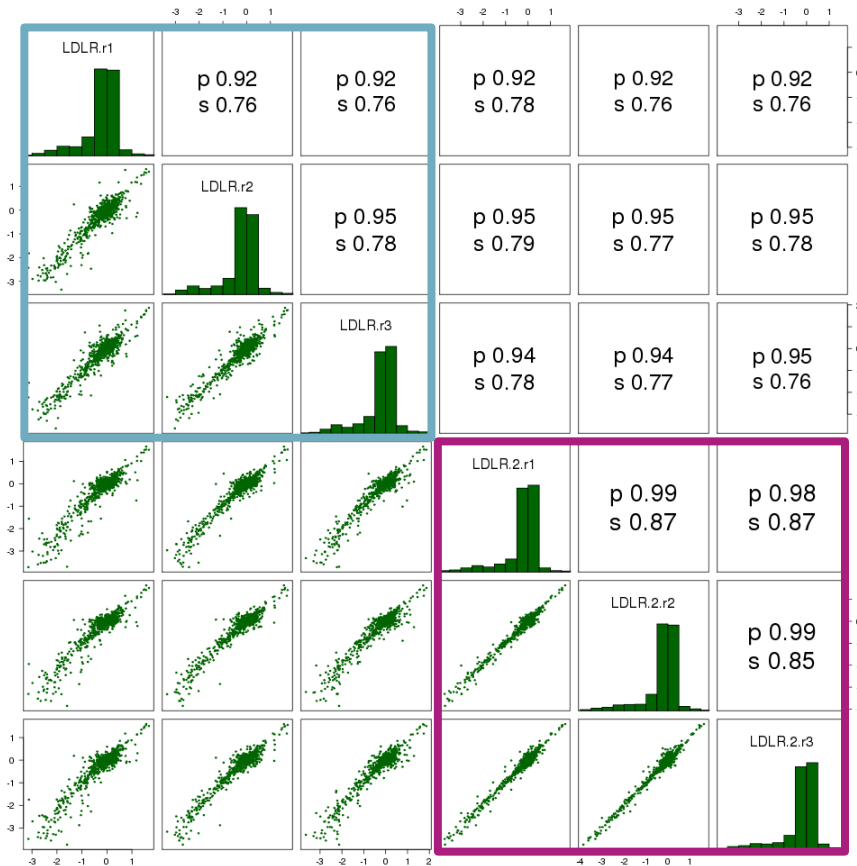
LDLR Full replicate I / Full replicate II



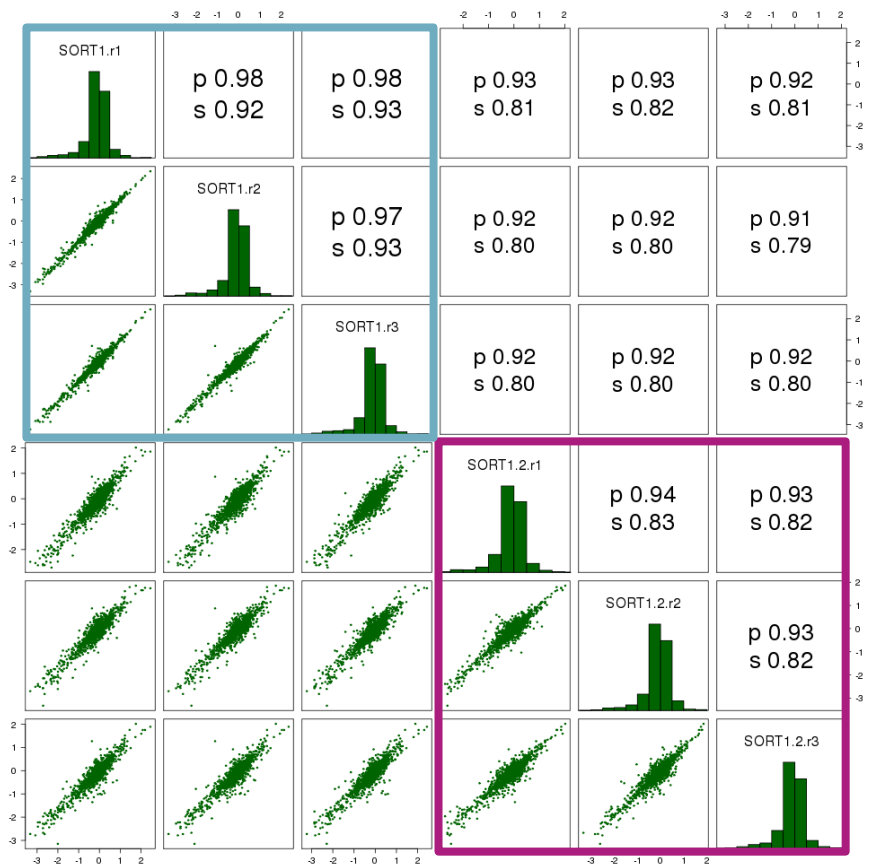
Combined replicates: p 0.97 / s 0.84

Reproducibility & experimental variation

LDLR Full replicate I / Full replicate II SORT1



Combined replicates: p 0.97 / s 0.84



Combined replicates: p 0.95 / s 0.85

Example: TERT promoter

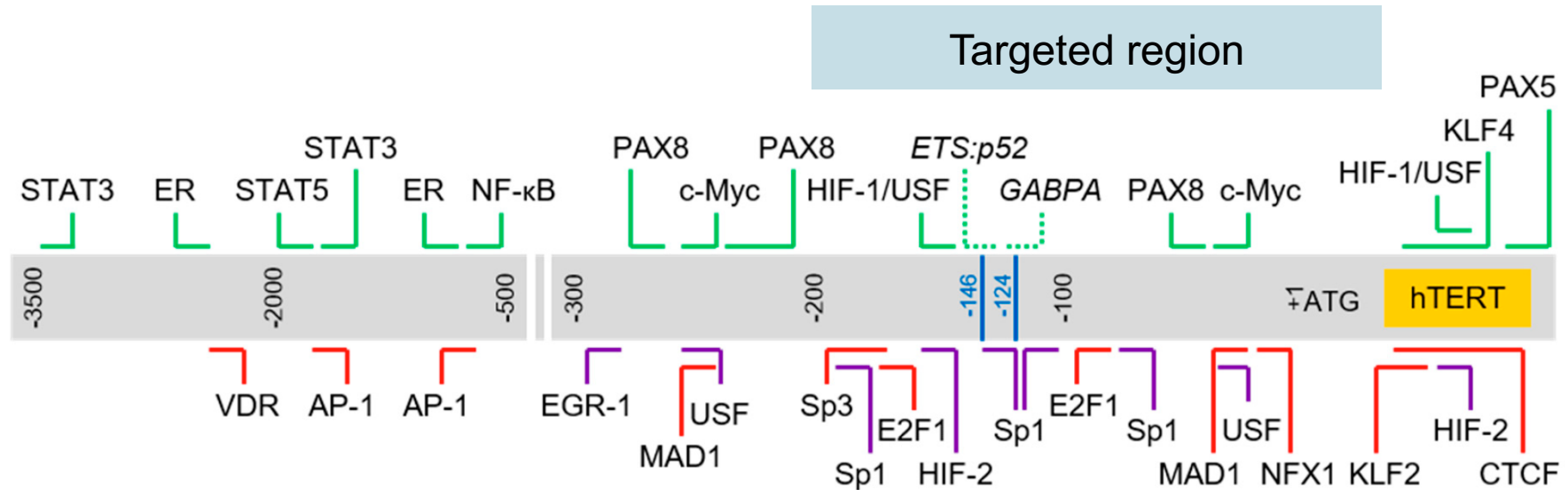
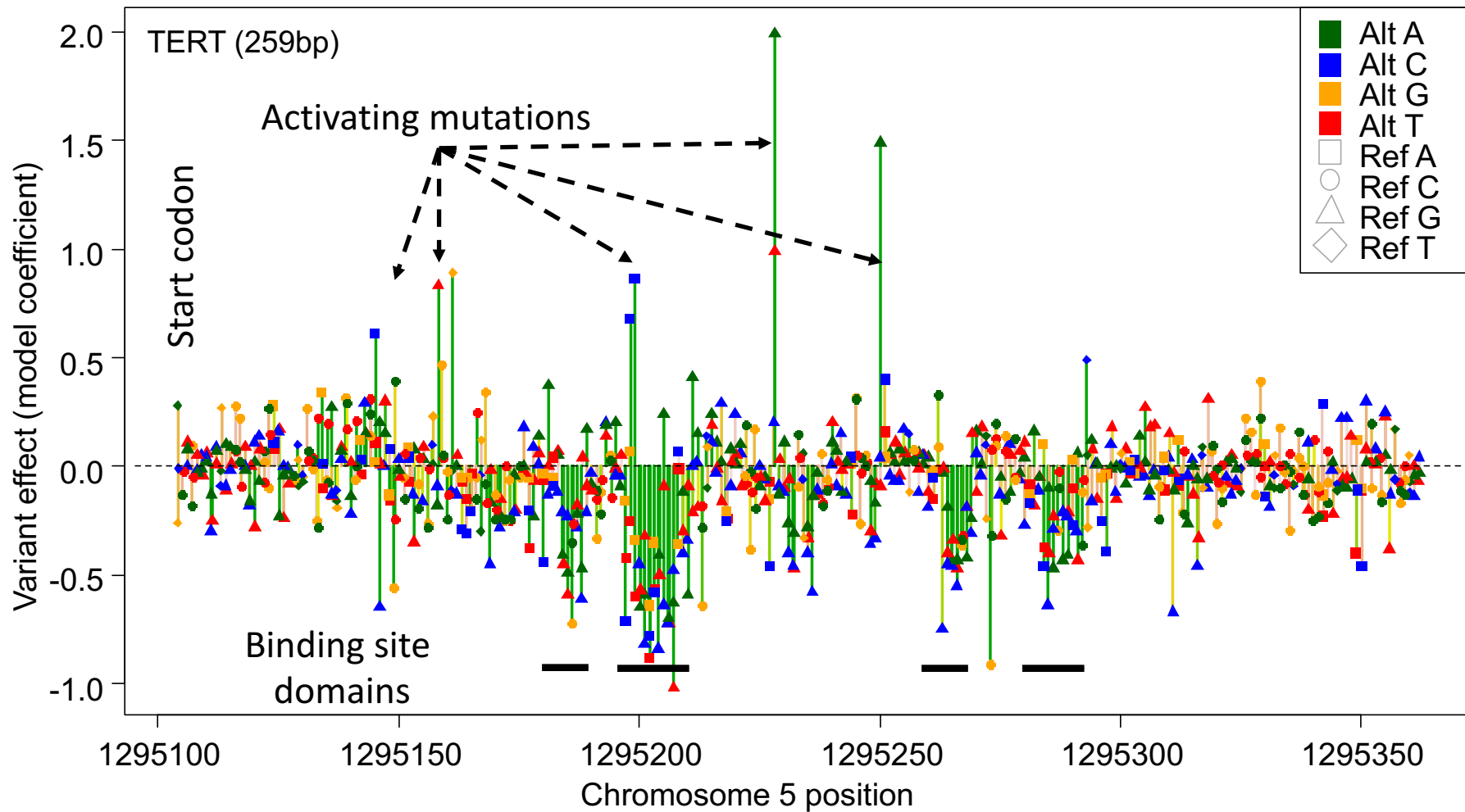


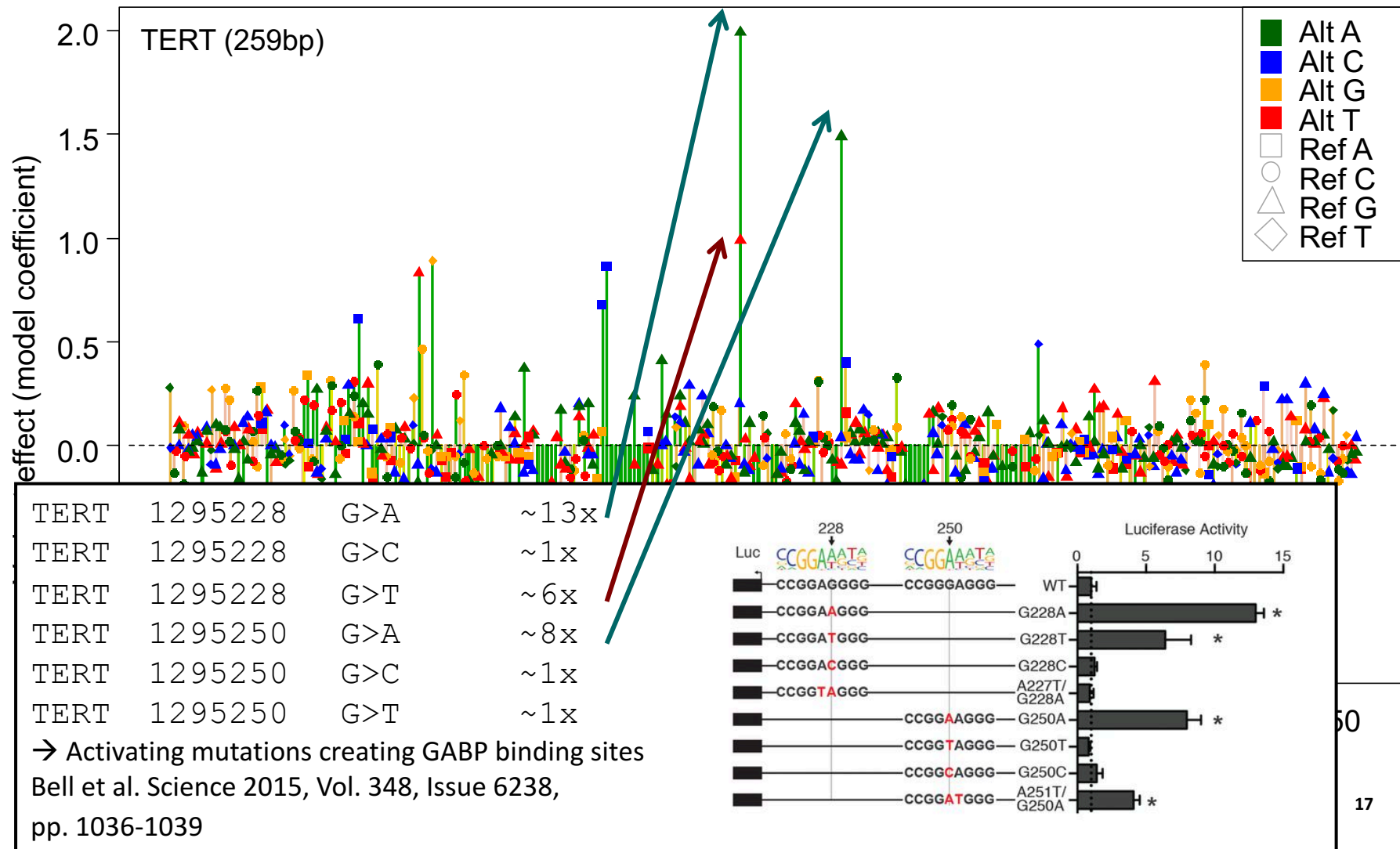
Figure 1. Schematic of transcription factor binding sites in human Telomerase Reverse Transcriptase (*hTERT*) promoter. Chromosomal sequence extending from 3.5 kb upstream and 150 bp downstream of *hTERT* translation start site (+1) is represented by the gray box. Horizontal lines above and below the box indicate approximate binding sites of respective transcription factors. Blue lines: hotspot promoter mutations (“-124” corresponds to C228T mutation; “-146” corresponds to C250T mutation); green: activator; red: repressor; purple: regulator with dual roles; dotted line: regulator bound to sites created by hotspot mutations.

Genes 2016, 7(8), 50; doi:10.3390/genes7080050

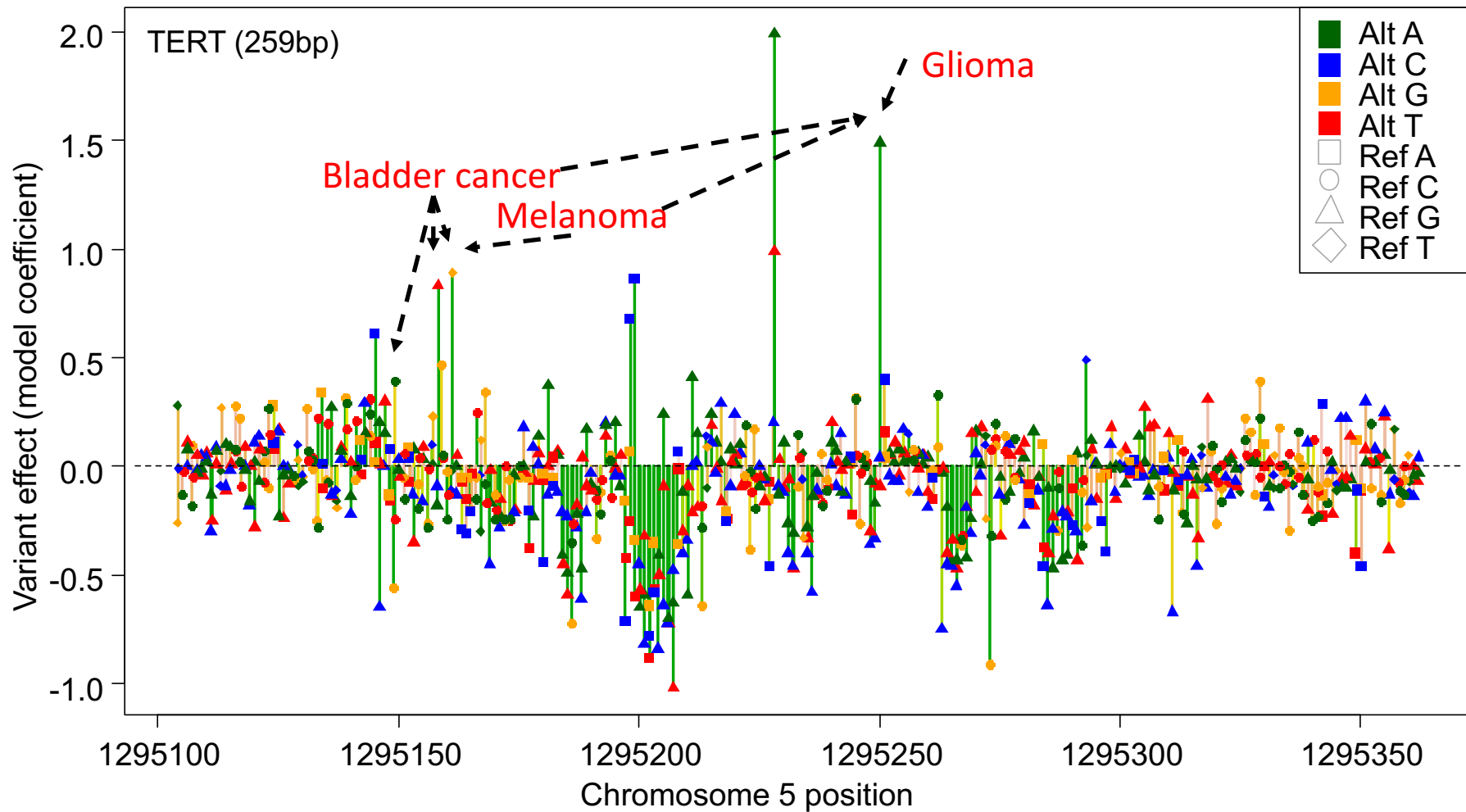
TERT promoter: HEK293T



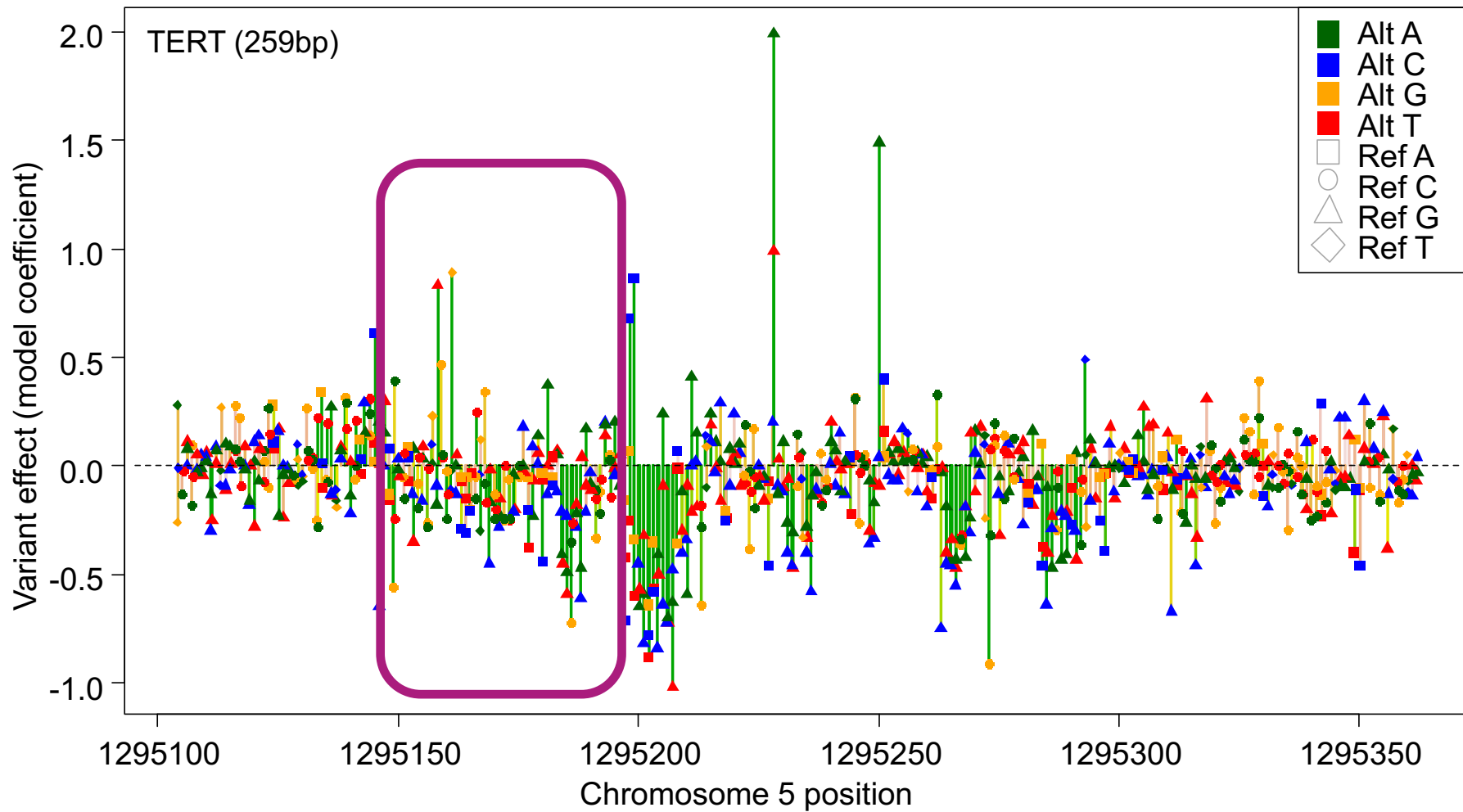
TERT promoter: HEK293T



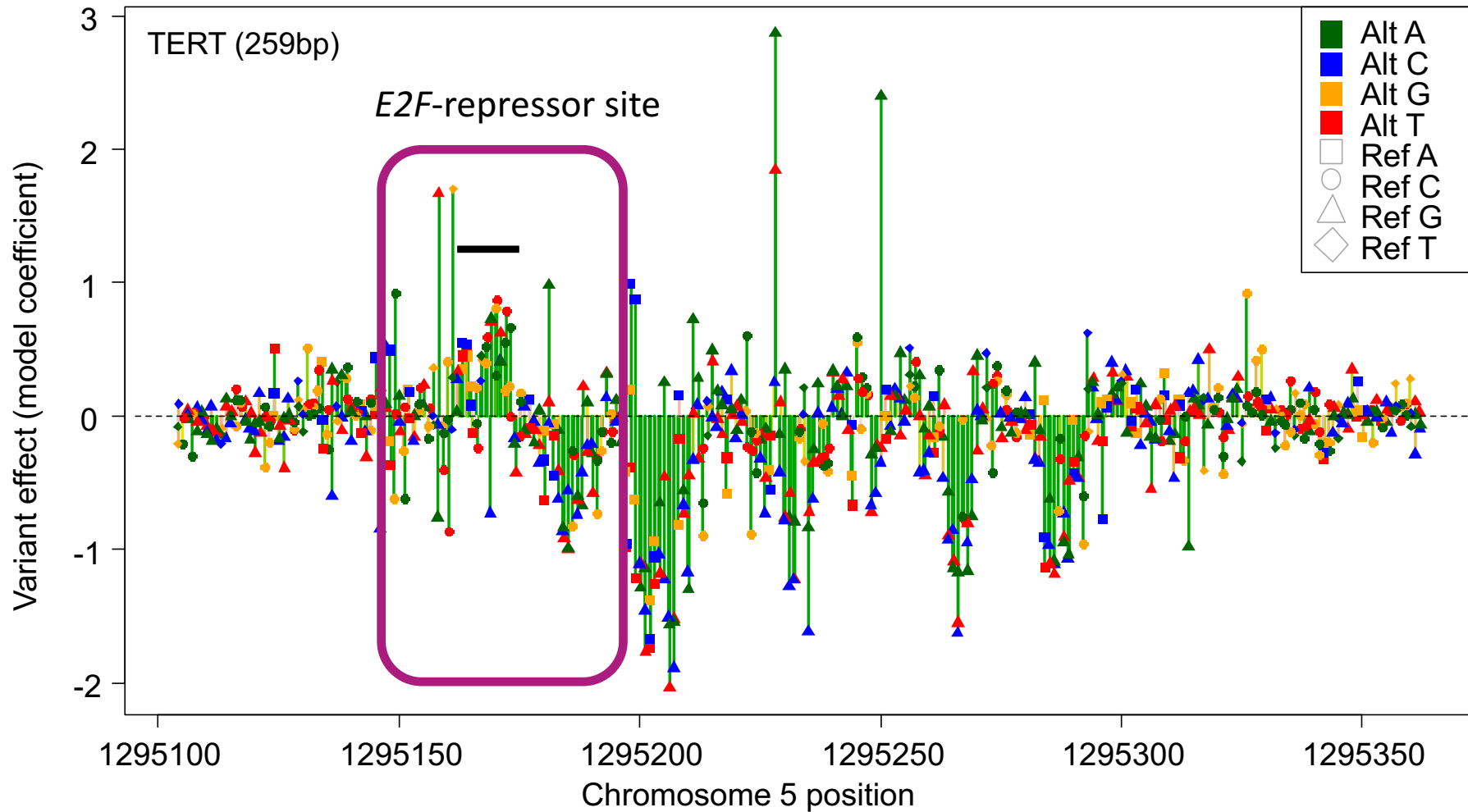
TERT promoter: HEK293T



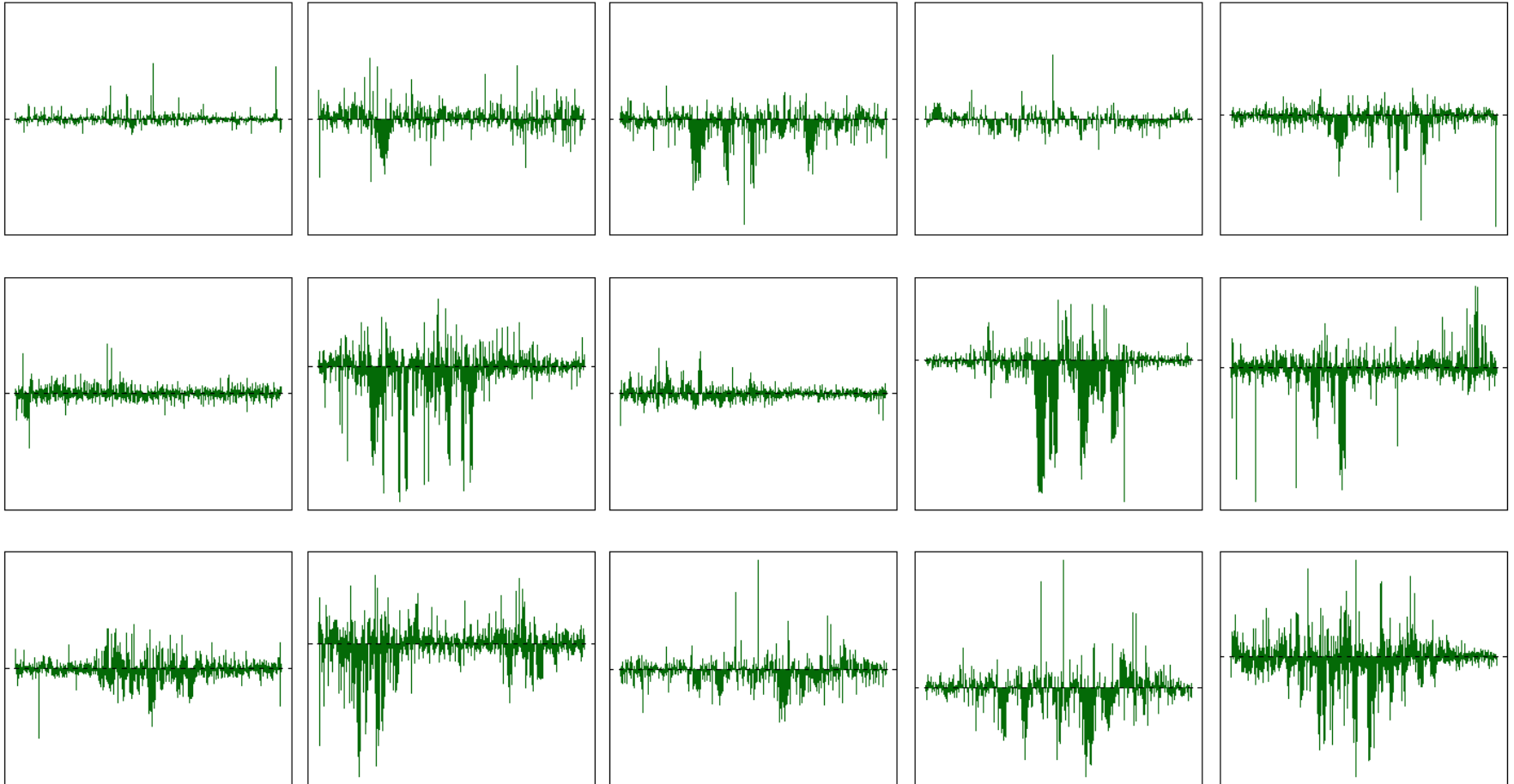
TERT promoter: HEK293T



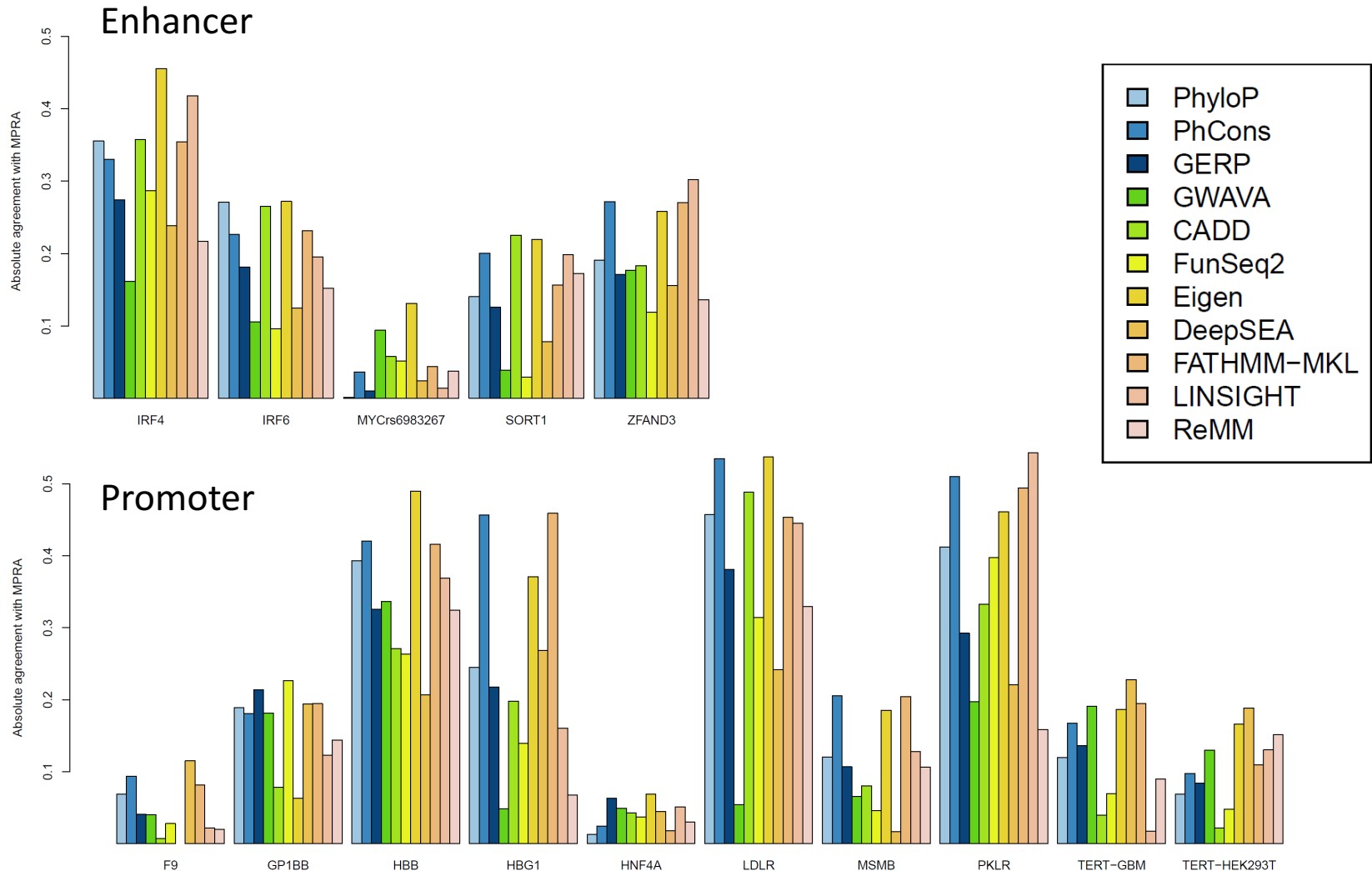
TERT promoter: glioblastoma cell-line



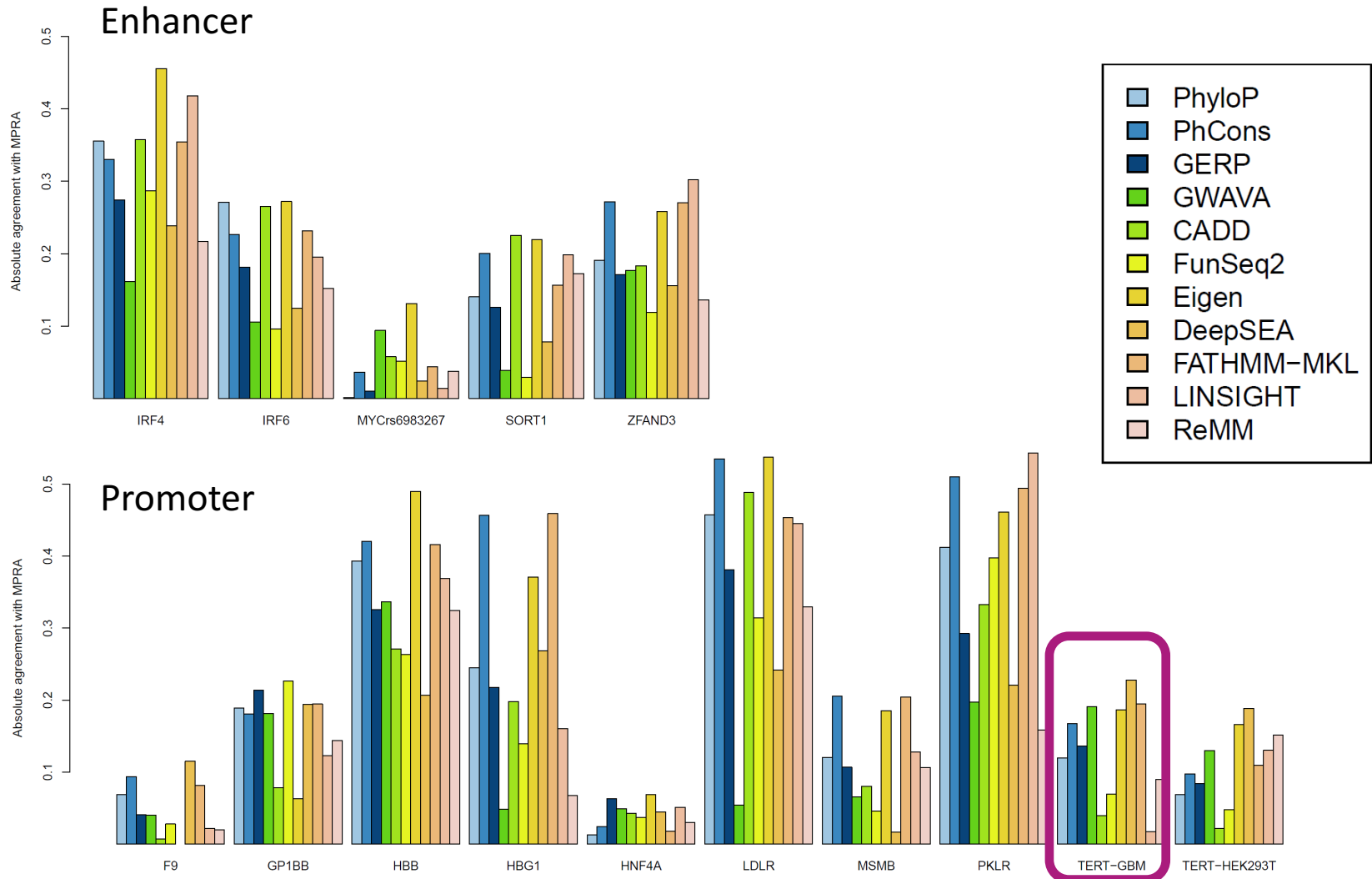
Saturation mutagenesis data sets part of CAGI 5



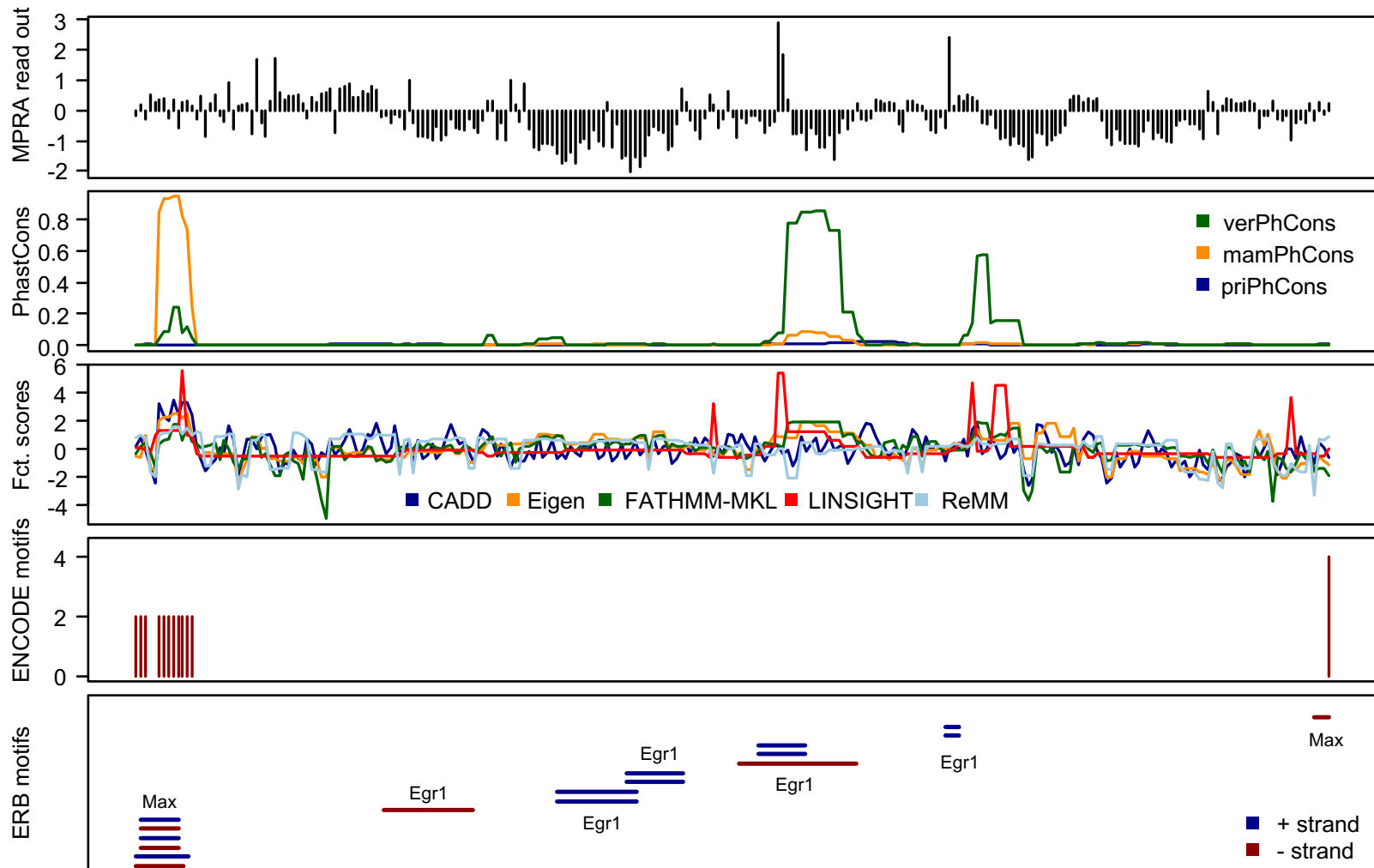
How do available variant scores perform?



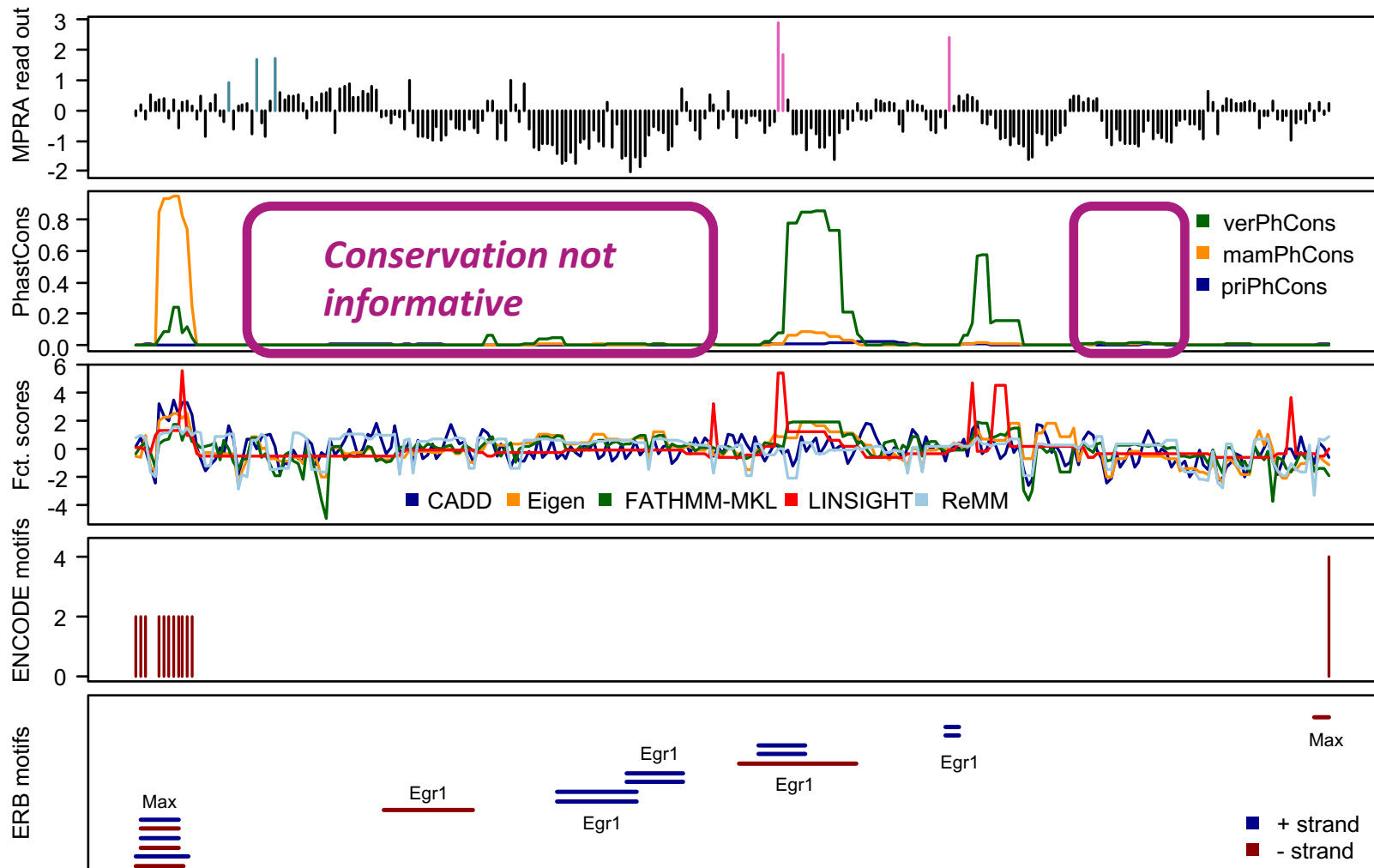
How do available variant scores perform?



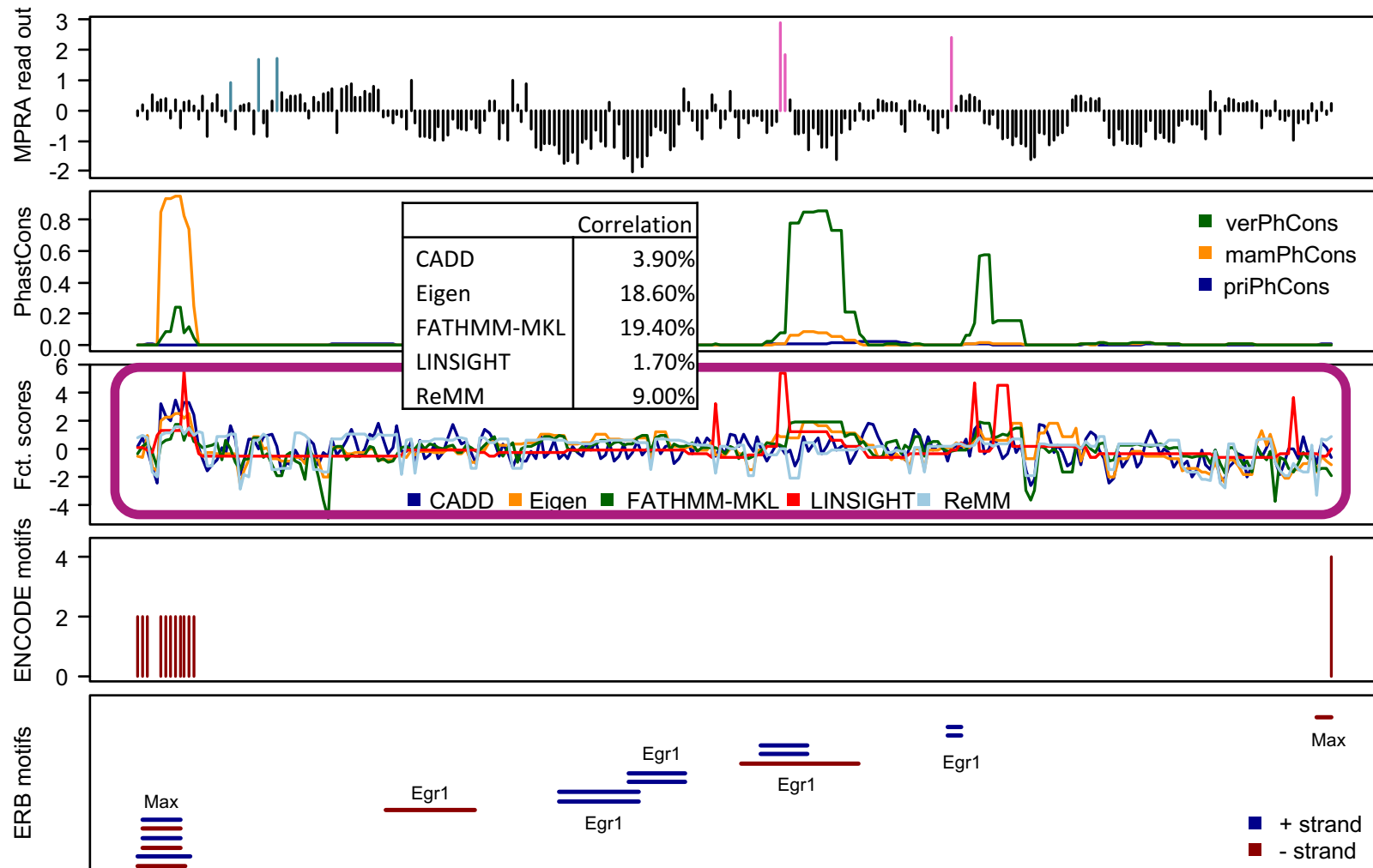
TERT promoter & annotation tracks



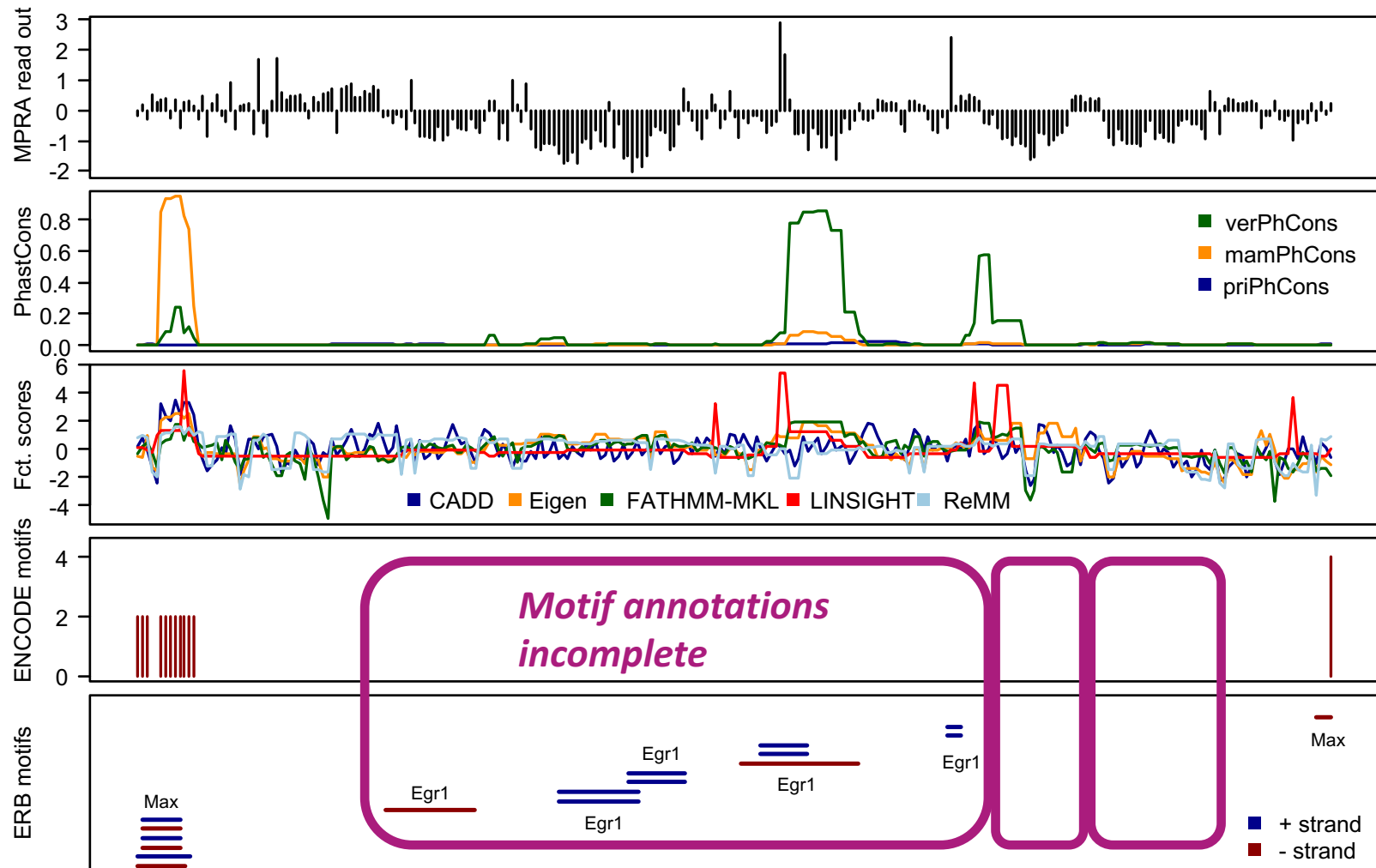
TERT promoter & annotation tracks



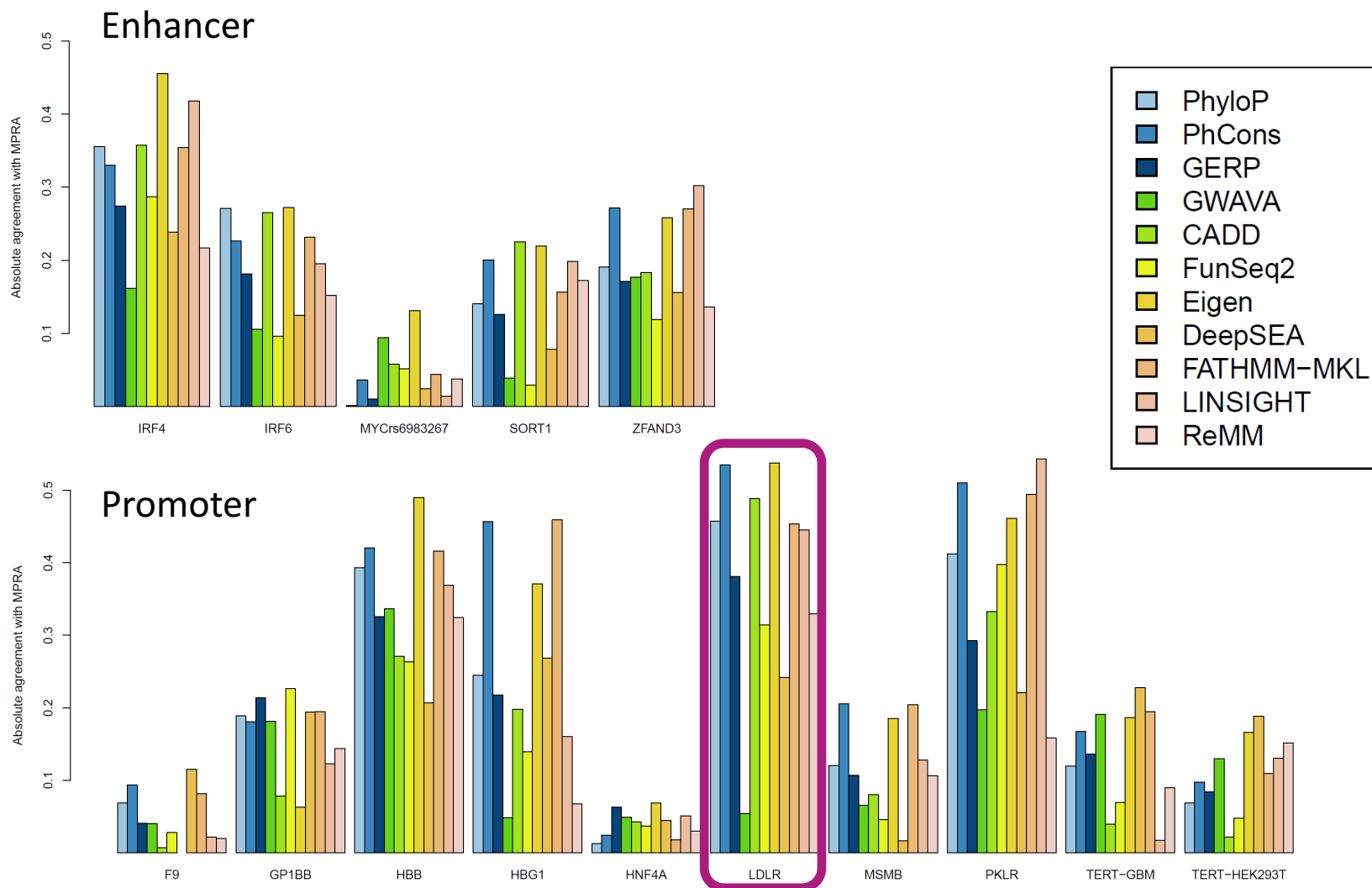
TERT promoter & annotation tracks



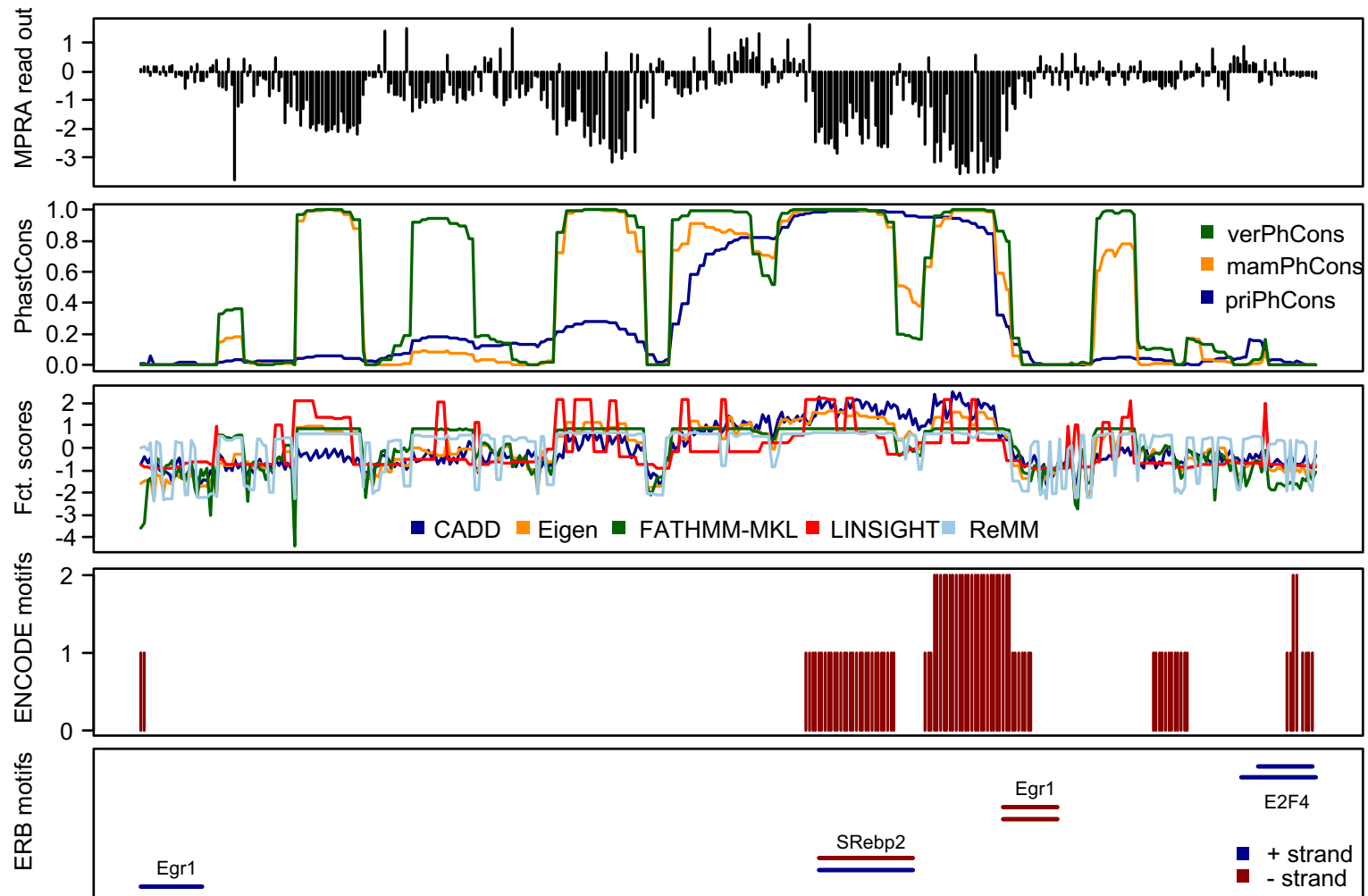
TERT promoter & annotation tracks



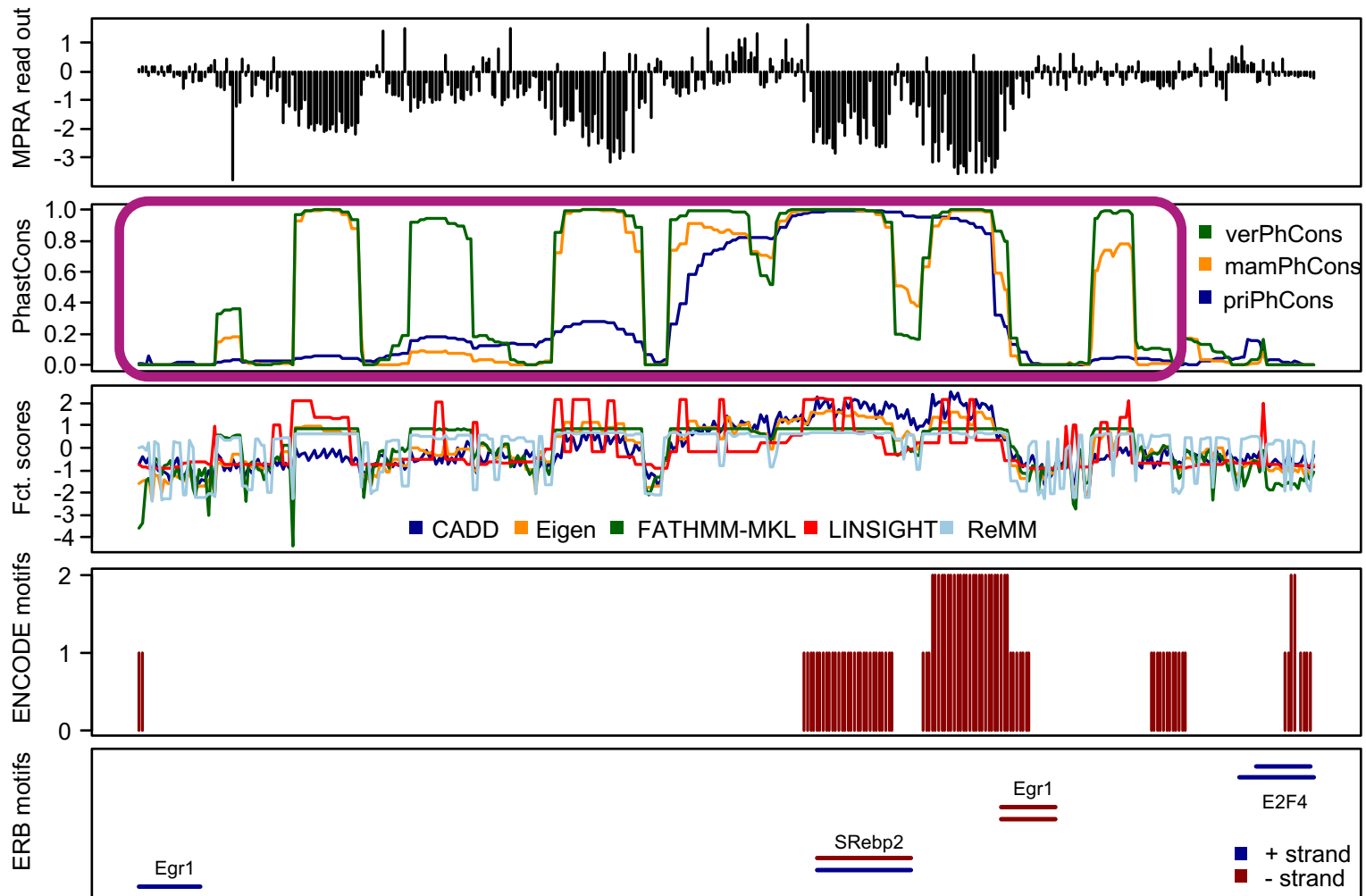
How do available variant scores perform?



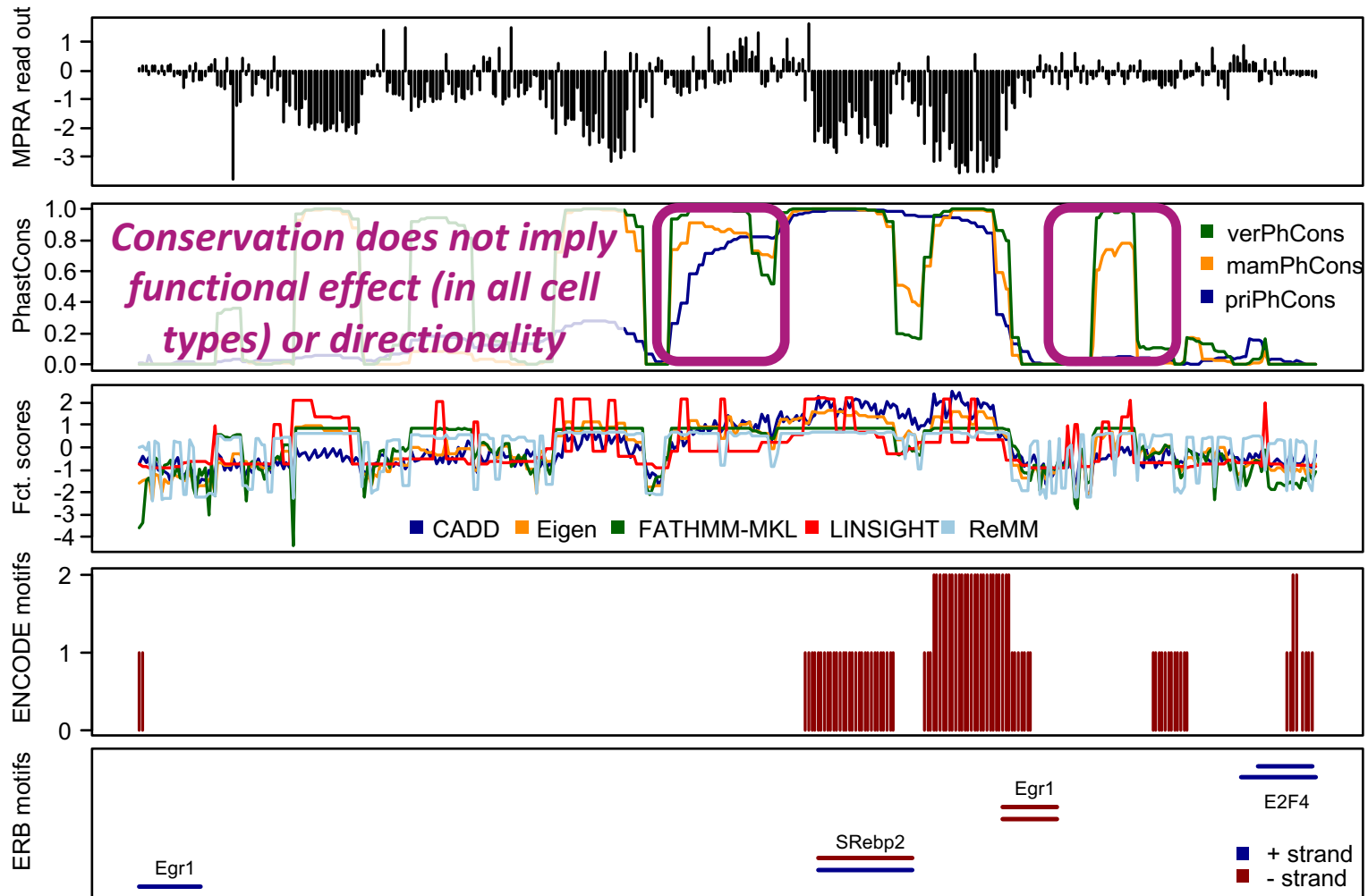
LDLR promoter



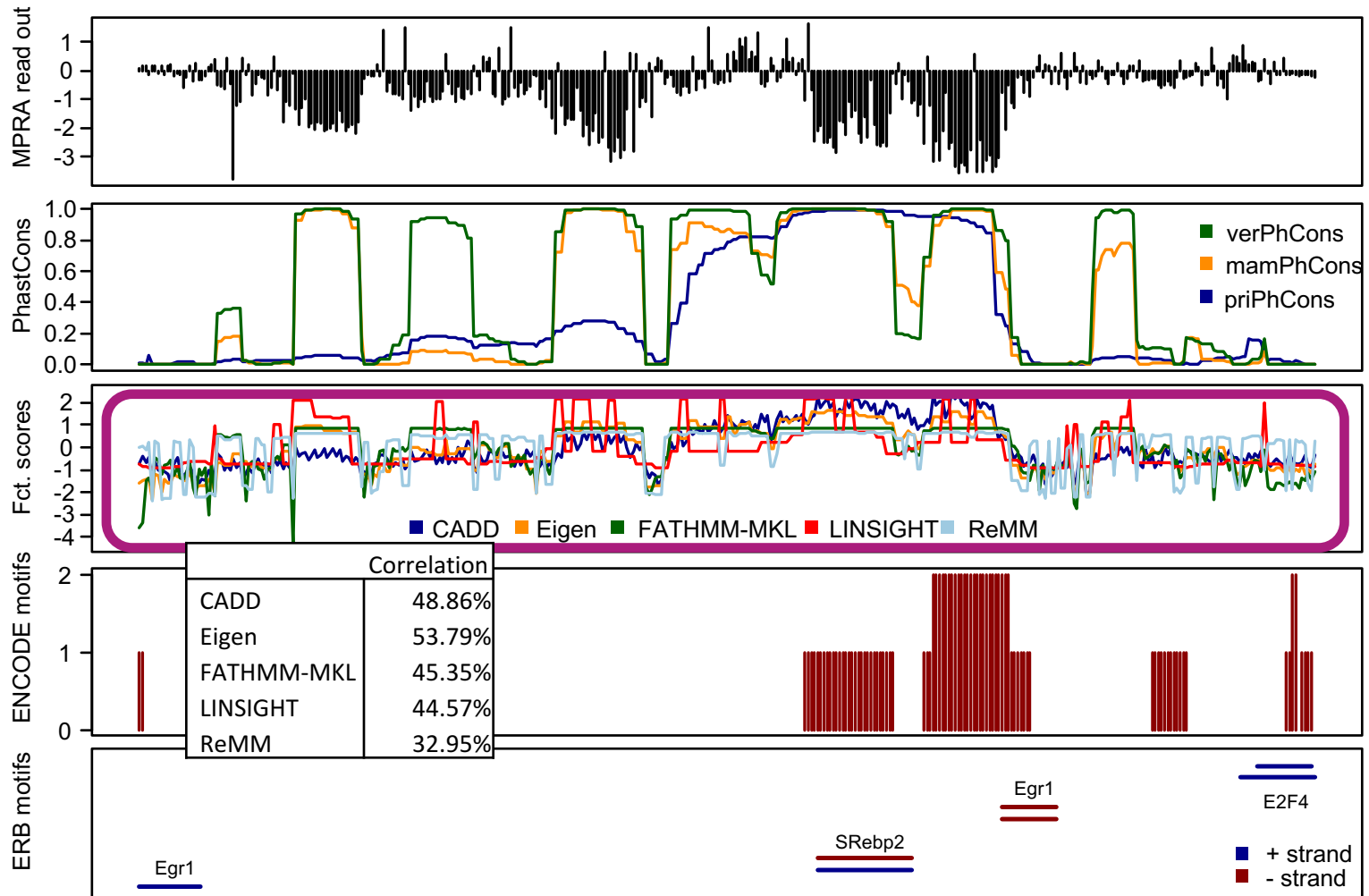
LDLR promoter



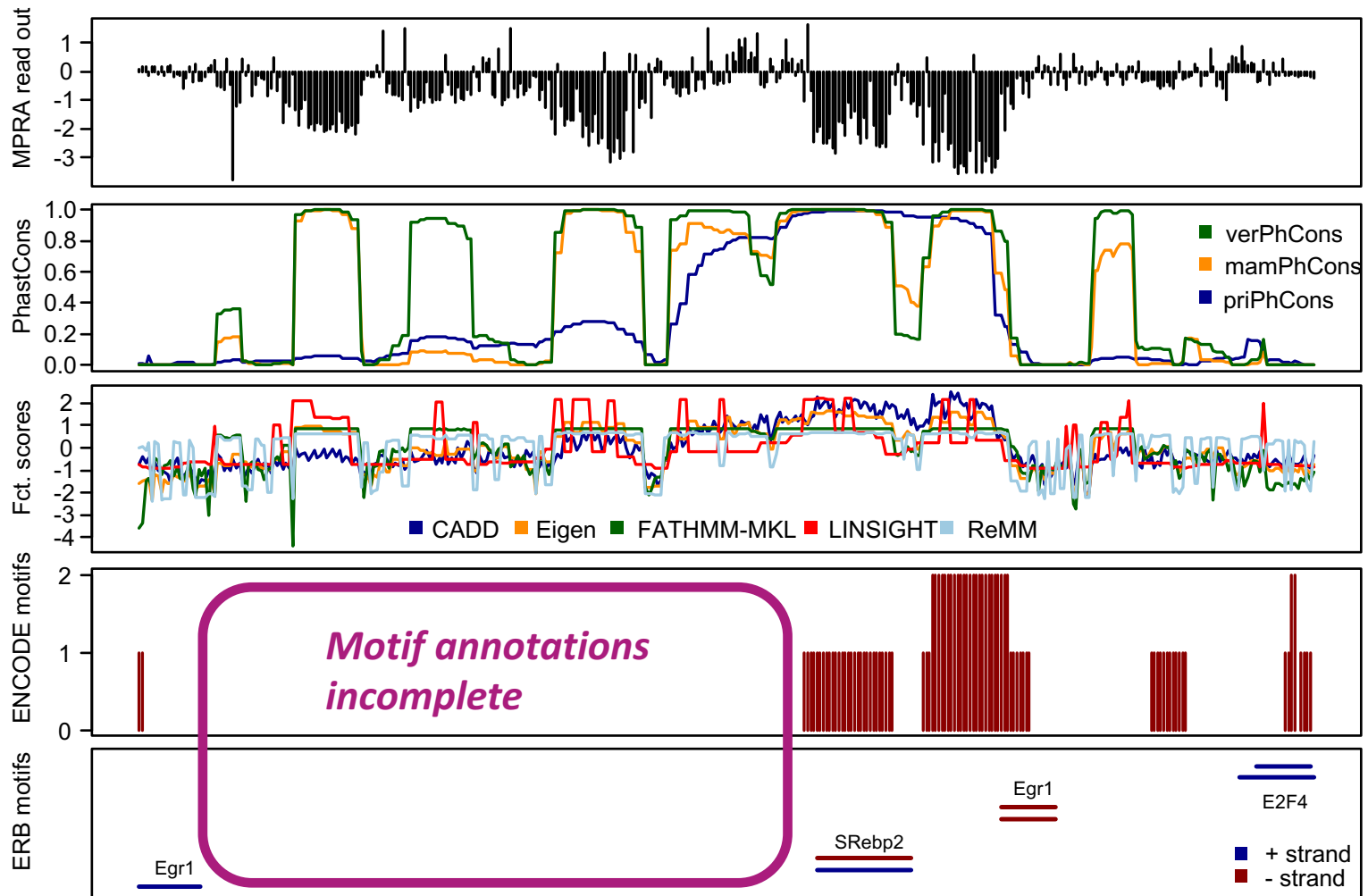
LDLR promoter



LDLR promoter



LDLR promoter



Summary

- Species conservation seems a major driver of available functional scores
- Conservation only provides sufficient information for a subset of elements
- Repressive effects align with available motif data, but predictions frequently incomplete
- Motif creation (activating mutations) currently not modeled

Eigen	0.276
PhCons	0.250
FATHMM-MKL	0.245
LINSIGHT	0.208
PhyloP	0.203
CADD	0.177
DeepSEA	0.156
GERP	0.175
FunSeq2	0.144
ReMM	0.142
GWAVA	0.125



HOW DO YOU PERFORM?



THANK YOU!

CONTACT

Max Schubach

max.schubach@bihealth.de

