Negative selection as an unbiased benchmark for evaluation of mis-splicing predictors

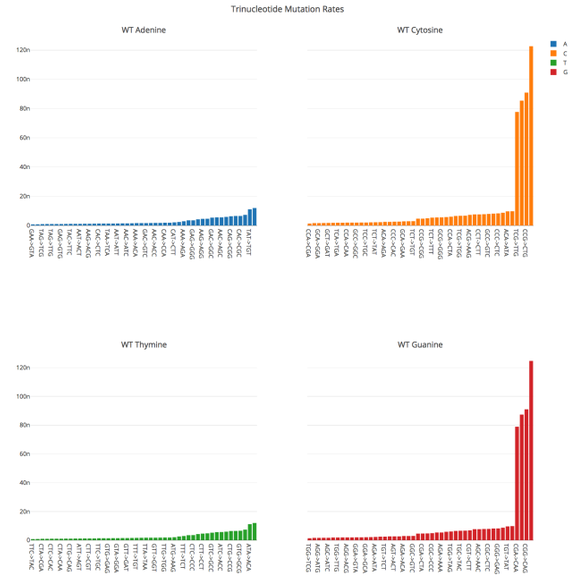
Philip Fradkin[1], Daniele Merico[1]

[1]Deep Genomics, 661 University Ave #480, Toronto, ON M5G 1M1

INTRODUCTION

The publicly available gnomAD dataset of over 120 thousand exomes allows for powerful negative selection analysis. Nucleotide mutation rates differ up to two orders of magnitude depending on the trinucleotide context and are thus important for analyzing selection. We learn the splicing consensus sequence from observed variants in haploinsufficient regions where heterozygous variants are under negative selection. Variants effect on splicing has previously been mostly studied in artificial systems before such as minigene reporters. gnomAD dataset allows for empirical investigation of the splicing code. Mis-splicing predictors in large have been trained on artificially constructed publicly available resources, making their validation tricky. In this project we utilize negative selection to benchmark mis-splicing predictors while controlling for the bias of trinucleotide mutation probabilities.

Trinucleotide mutation rates



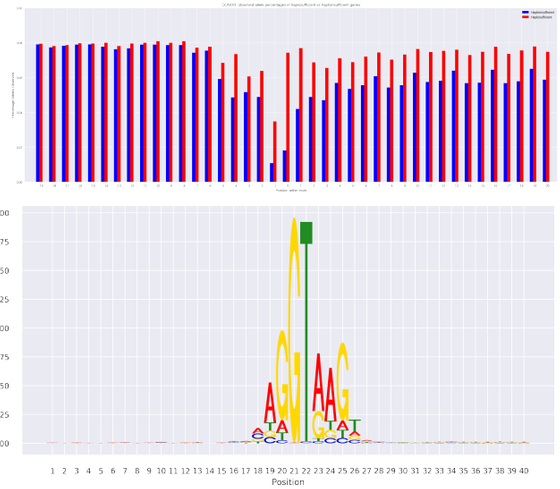
AIM

1. Examine the effect of trinucleotide mutation rate on negative selection.
2. Analyze splicing code using empirically observed variants surrounding the consensus region.
3. Provide an unbiased evaluation of mis-splicing predictors using negative selection in the intronic and exonic consensus regions.

METHODS

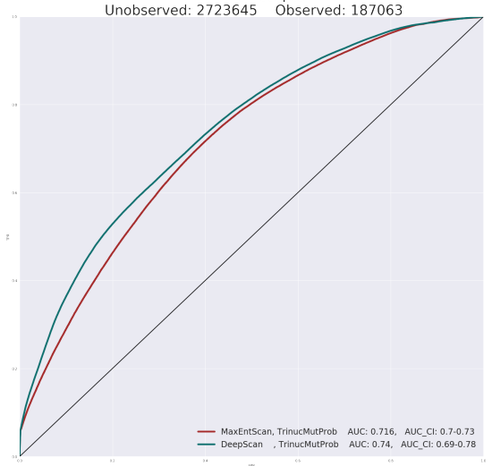
1. Identify all variants in RefSeq transcriptome
2. Subset to haploinsufficient genes
3. Treat all variants **unobserved** in gnomAD as causing mis-splicing (positives)
4. Treat all variants **observed** in gnomAD not causing mis-splicing (negatives)
5. Score variants using MaxEntScan and DeepScan in splicing consensus regions [e1-e3], [i3-i6], [i3-i20]
6. Fit a logistic regression with predictors and trinucleotide mutation probability

Donor observed constraint

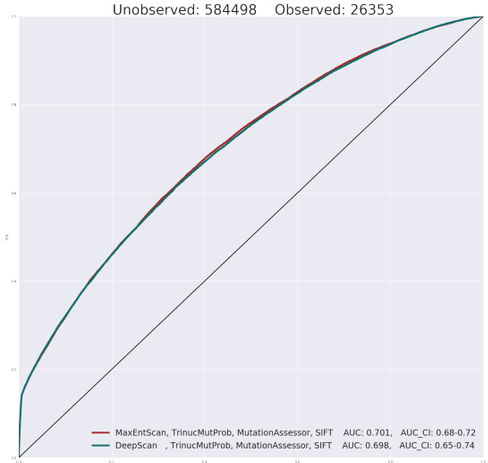


Acceptor observed constraint

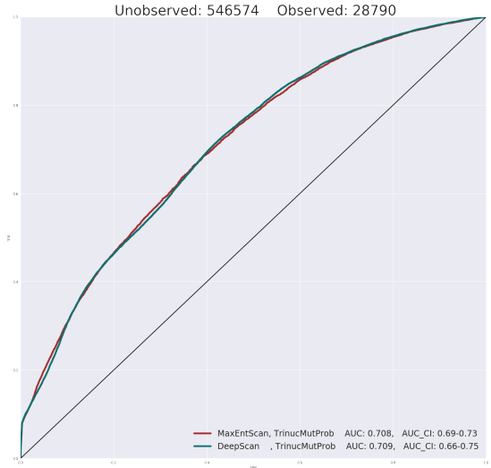
Intronic Acceptor
Unobserved: 2723645 Observed: 187063



Exonic Consensus
Unobserved: 584498 Observed: 26353



Intronic Donor
Unobserved: 546574 Observed: 28790



RESULTS

We empirically investigate the splicing consensus code, using genomic sequence in haploinsufficient genes. We then use observed gnomAD variants to evaluate mis-splicing predictors in the consensus sequence (MaxEntScan defined) excluding the intronic dinucleotide. All unobserved variants are assumed to cause mis-splicing. Although some variants will be unobserved due to stochastic reasons this assumption allows for an unbiased relative ranking of molecular phenotype predictors. We fit a logistic model using predictors and trinucleotide mutations probability as features. We demonstrate that DeepScan performs significantly better than MaxEntScan in the intronic acceptor region as defined by MaxEntScan ($p = 2.2e\text{-}165$ value two sided Wilcoxon rank test). Molecular phenotype predictor comparison in the exonic region fails due to an insignificant percentage of negative selection associated with mis-splicing. We control for deleterious amino acid substitution effects by adding Mutation Assessor and SIFT as covariates.

CONCLUSIONS

Observed negative selection can function as a ranking tool for molecular phenotype predictors if trinucleotide mutation probability is accounted for. While less powerful than traditional mutagenesis approaches, this evaluation allows for an unbiased performance ranking. Absolute performance of the predictors is hard to estimate since we are unable to determine the cause of unobserved variants: due to stochasticity or negative selection. As public genomic datasets become bigger the power of this approach increases since our negatives become enriched with variants under negative selection.

REFRENCES

1 Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *Nature* 536 (7616): 285–91.

2 Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. 2007. "NCBI Reference Sequences (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins." *Nucleic Acids Research* 35 (Database issue): D61–65.

3 Sunyaev, Shamil, Fyodor A. Kondrashov, Peer Bork, and Vasily Ramensky. 2003. "Impact of Selection, Mutation Rate and Genetic Drift on Human Genetic Variation." *Human Molecular Genetics* 12 (24): 3325–30.

4 Yeo, Gene, and Christopher B. Burge. 2004. "Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 11 (2-3): 377–94..

ACKNOWLEDGEMENTS

Thank you to Deep Genomics for sponsoring this research. As well as Amit Deshwar, Alice Gao, Mark Sun, Michael Weinberg and Brendan Frey for the helpful input.