Assessment of the performance of splicing predictors at non-canonical intronic sites and implications for variant classification

David H. Tran, Peter Kang, Rebecca Mar-Heyming, Saurav Guha

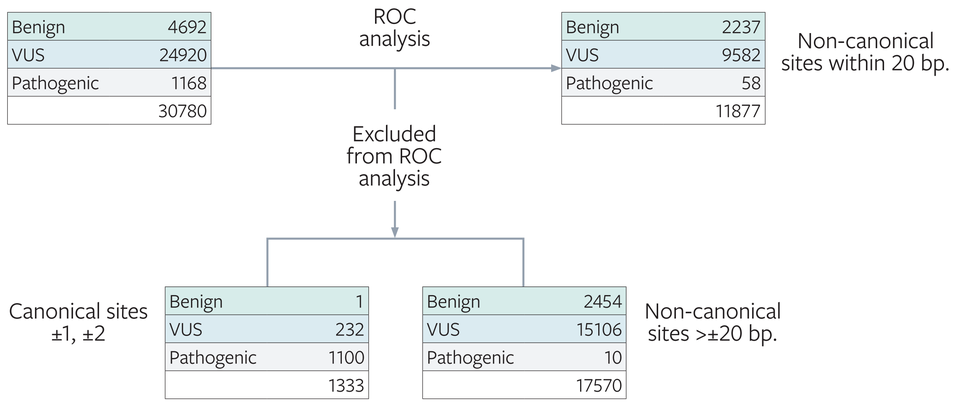
South San Francisco, California

Introduction

Genetic variants that disrupt or alter normal splicing are important factors underlying many genetic disorders. The appropriate clinical classification of these variants represents a challenge that could be assisted by accurate in silico predictors and comprehensive population-genetic information. Here we leveraged the Counsyl variant database to evaluate the general features of intronic-variant classification. We compared the population frequencies of intronic SNPs with variant classification. In addition, we assessed the performance of in silico predictors on the classification of intronic SNPs (focusing on variants within 20 bp. of the splice junction but outside the "canonical" ±1, ±2 positions).

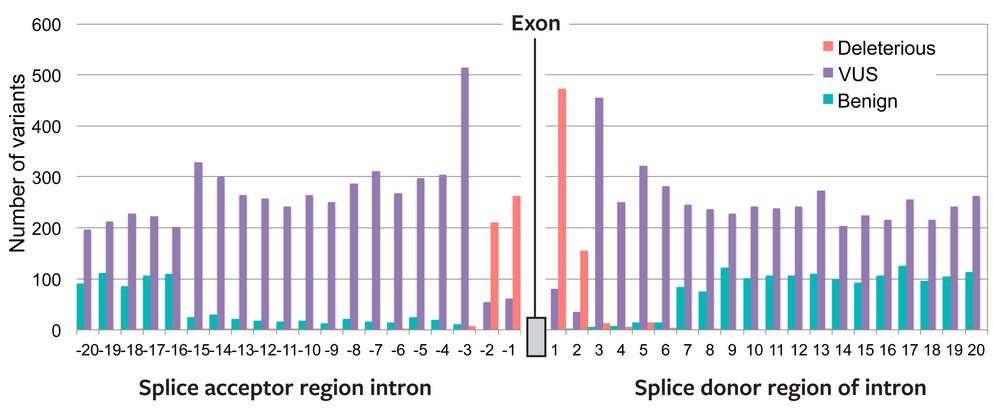
Overview of dataset

30,780 curated, intronic SNP alleles comprised our initial dataset, pulled from ~500,000 samples tested on Counsyl's Foresight Carrier Screen (224 genes). The final data set included 11,877 alleles at non-canonical sites within 20 bp. of the splice junction.



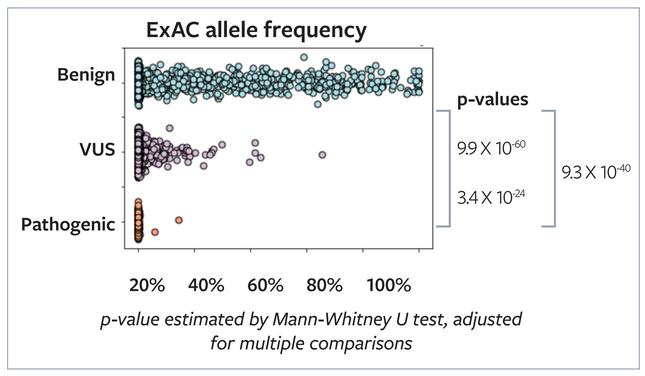
58 (4.9%) of these alleles were classified pathogenic while 2237 (18.8%) were classified benign.

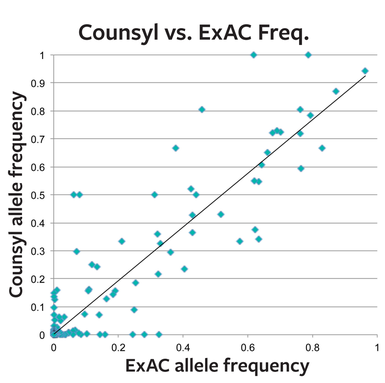
Classification by intronic position



Classification of variants at positions within 20 bp. of the junction. There is a general gradient of pathogenic-VUS-benign classifications the farther away the site is from the junction. Interestingly there is an approximately 1.5- to 2-fold increase in the number of variants at positions -3, +1, and +3.

ExAC allele frequencies in the sample cohort





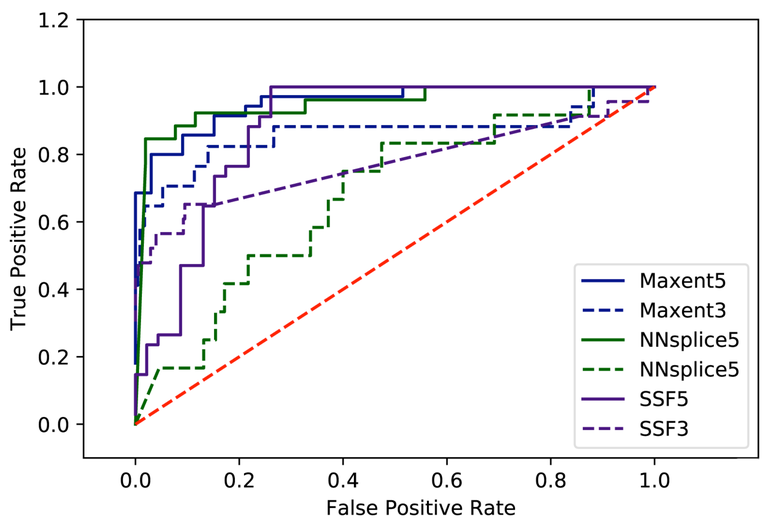
15,192 alleles were reported in ExAC (Exome Aggregation Consortium), note these include intronic positions greater than 20 bp from the junctions. Population allele frequencies for intronic SNPs showed a significant stratification between classifications.

A random sample of Counsyl allele frequencies (n=752) did not show a substantial deviation from ExAC allele frequencies. These data suggest that our sample cohort was not significantly divergent from the general population.

Specificity/sensitivity analyses (by receiver operator characteristics)

We evaluated the performance of three in silico splice predictors: MaxEnt, NNSplice, and SSF on non-canonical variants that had been classified by literature and database review as either pathogenic or benign (excluding VUS). These analyses determine whether these tools accurately predicted the clinical impact of these variants along with splicing defect. In other words, did in silico splicing predictions correlate well with a clinical assessment of pathogenicity?

Comparison of ROC curves for each tool, split by splice region



Summary of performance analyses

				NUMBER OF VARIANTS			
Tools	**Optimum threshold**	**Sensitivity**	**Specificity**	**Pathogenic**	**Benign**	**Sequence windows**	**AUC**
maxent5	63.5%	85.7%	90.9%	35	33	+3 to +6	0.95
nnsplice5	84.3%	92.3%	88.5%	26	52	+3 to +8	0.95
ssf5	94.3%	88.2%	78.3%	34	46	+3 to +7	0.88
maxent3	87.3%	82.4%	86.0%	17	229	-3 to -20	0.86
ssf3	96.9%	65.2%	90.5%	17	147	-3 to -14	0.77
nnsplice3	99.6%	75.0%	60.0%	12	175	-3 to -20	0.68

The optimum threshold is defined as the threshold of variant score (estimated strength of splice variant) divided by the reference score (estimated strength of the reference site), which provides the maximal AUC score. For example, a 63.5% threshold suggests that classifying variants based on scores below 63.5% of the reference score optimizes the sensitivity and specificity of the prediction. The sensitivity is the rate of true positives and the specificity is the rate of true negatives when the optimum threshold is used in prediction of pathogenicity.

Sequence windows are the effective distance the splicing tool has for predicting splice defects at non-canonical sites. The AUC (or area under the curve) are the areas under the curve in the chart above, ranging from 0 to 1, with 1.0 achieving the best possible performance (sensitivity and specificity of 100%).

Conclusion

A combination of population allele frequency and in silico prediction can provide some useful insights into potential clinical impact of intronic SNPs. The majority of pathogenic variants reside at the canonical positions while a relatively greater number of VUS variants lie at the acceptor region. The distribution of variant frequencies showed intergroup differences between the three classifications (benign, VUS, and pathogenic). The three tools tested did not perform as well for non-canonical sites acceptor sites.The best performing tools, overall, were MaxEnt and NNsplice.