

Supplemental

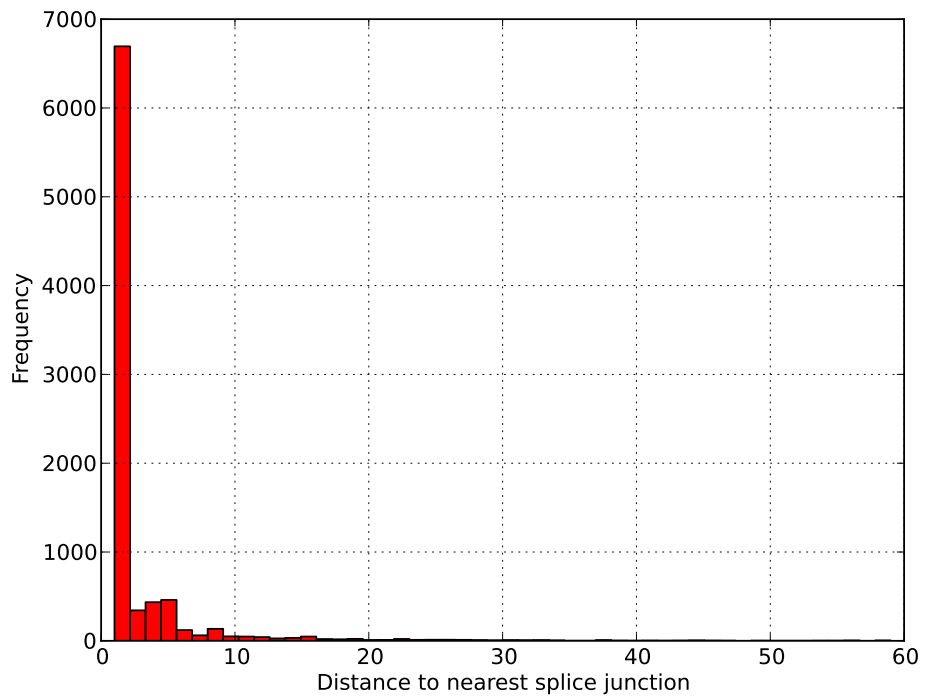


Figure 1: Distribution of distance to nearest splice junction of known HGMD mutations.

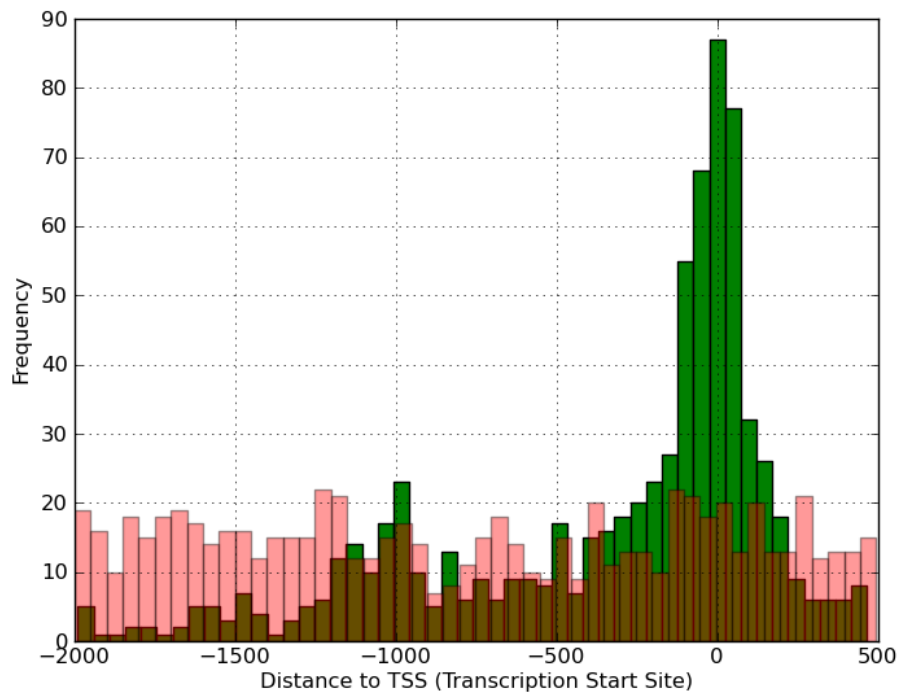


Figure 2: Comparison of distance to TSS. Green is the distribution of the distance to transcription start site of HGMD annotated variants, and red shows the distribution of the randomly selected background dataset.

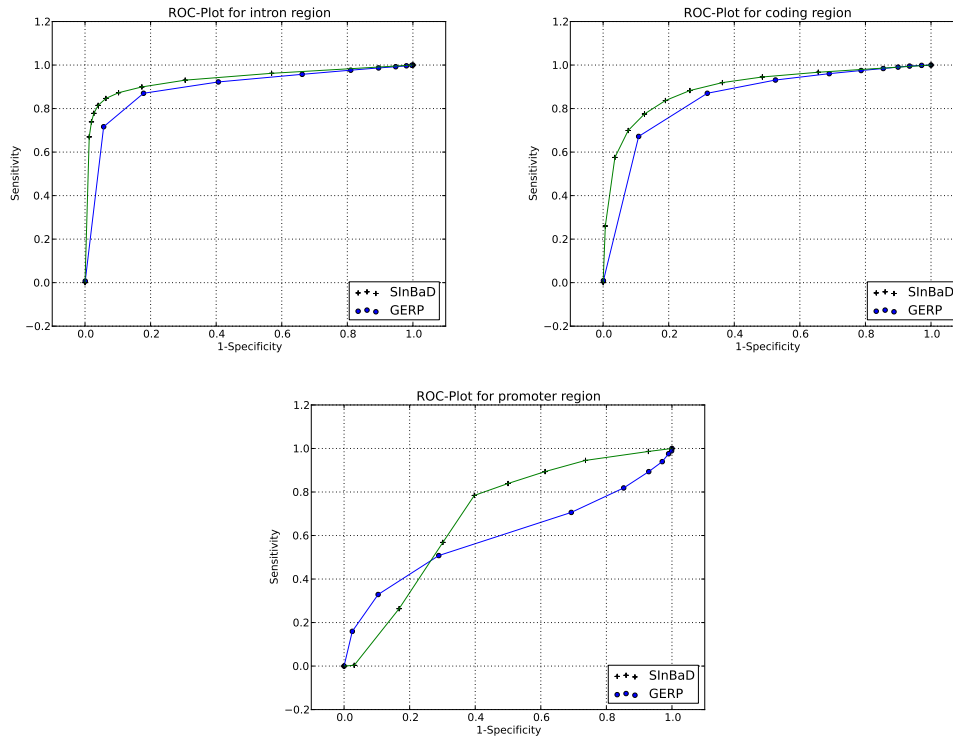


Figure 3: Specificity/Sensitivity comparison between SInBaD and GERP across the different gene regions.

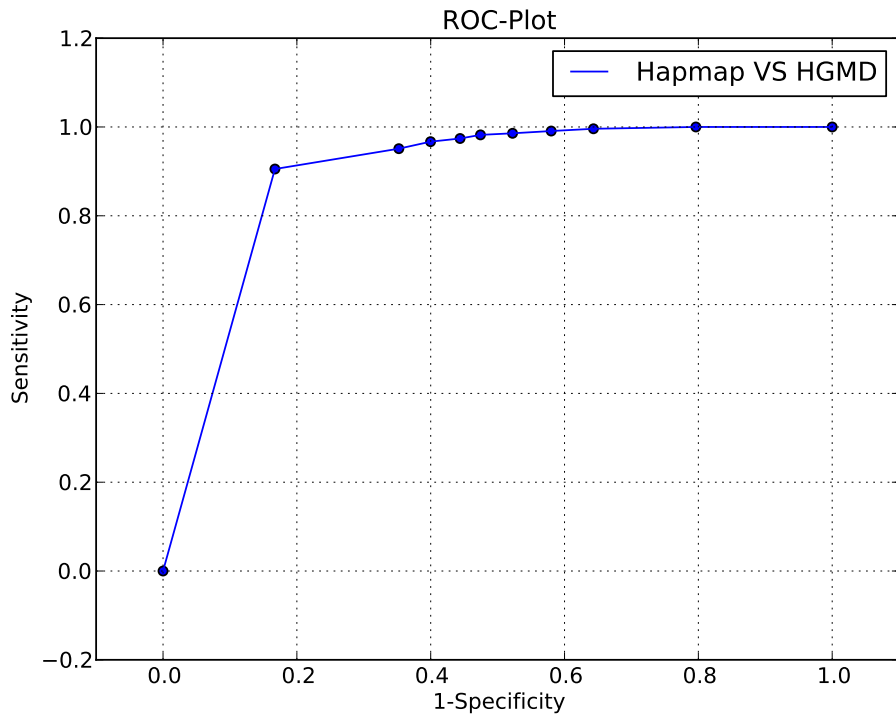


Figure 4: Comparison of Sensitivity/Specificity trade-off using 2,000 HapMap variants as supposedly non-functional variants and selecting 2,000 random disease mutations(50% coding and 50% intron) taken from HGMD. The model reasonably distinguishes HapMap variants from actual disease variants.

	Accuracy	Specificity	Sensitivity	Size
-500bp to +500bp(conservation)	56.57%	62.14%	51.20%	1,408
-2,000bp to +500bp(conservation)	59.37%	61.16%	57.6%	1,751
-500bp to +500bp(conservation + distance)	58.07%	58.09%	58.22%	1,408
-2,000bp to +500bp(conservation + distance)	71.31%	80.54%	62.21%	1,751

Table 1: 10 fold cross-validation results comparing training data in promoter region, selected by different window definitions around the transcription start site.