

## Abstract

Predicting and proper ranking of splice sites (SS) is a challenging problem in bioinformatics and machine learning communities. Proposed method of donor and acceptor SSs prediction is based on counting oligonucleotide frequencies for splice and splice-like signals. Based on bayesian principle SS sensors were built. We demonstrate advantage of our proposed sensor design compared with existing sensors and tools. In particular, our donor sensor outperforms Maximum Entropy Sensor for several representative test sets of genes when compared on Receiver Operating Characteristic (ROC) curve. We represent combinatorial interaction of SSs and related factors with Logarithm Of oDds (LOD) weight matrices. Based on factor interactions we were able to substantially improve splice signals prediction quality and rank SSs better than SpliceView, GeneSplicer, NNSplice and Genio tools. Proposed method is used in our new splicing simulator SpliceScan. Program, learning set and test results are available at <http://bioinformatics.ist.unomaha.edu/~achurban/>.

## Introduction

The precise removal of introns from pre-messenger RNAs (pre-mRNAs) by splicing is a critical step in expression of most metazoan genes. The process requires accurate recognition and pairing of 5' and 3' SSs by the splicing machinery. Inappropriate splicing of a gene may result into the translation of a non-functional protein.

Weakly conserved splice signals are necessary, but not sufficient for the exact recognition of the exons. Frequently degenerate donor, acceptor, polypyrimidine tract and the branch point motifs provide insufficient information for the exact SSs detection.

Correct prediction of SSs appear to be the key ingredient to successful *ab initio* gene annotation, since dynamic programming procedures have to see all the exon/intron boundaries in order to find the optimal solution [1]. The most sensitive sensor design predicting the least amount of false positives is preferable. Another good feature of a SS sensor is ability to rank predicted SSs, i.e. assign certain score characterizing importance or strength of a putative site of splicing.

Figures 1,2 show SS consensus for both 5' and 3' exonic ends. Human genome has plenty of motifs of unknown functionality with structure very similar to just mentioned consensus. That sites are called *splice-like sites* and they outnumber the real sites by several orders of magnitude.

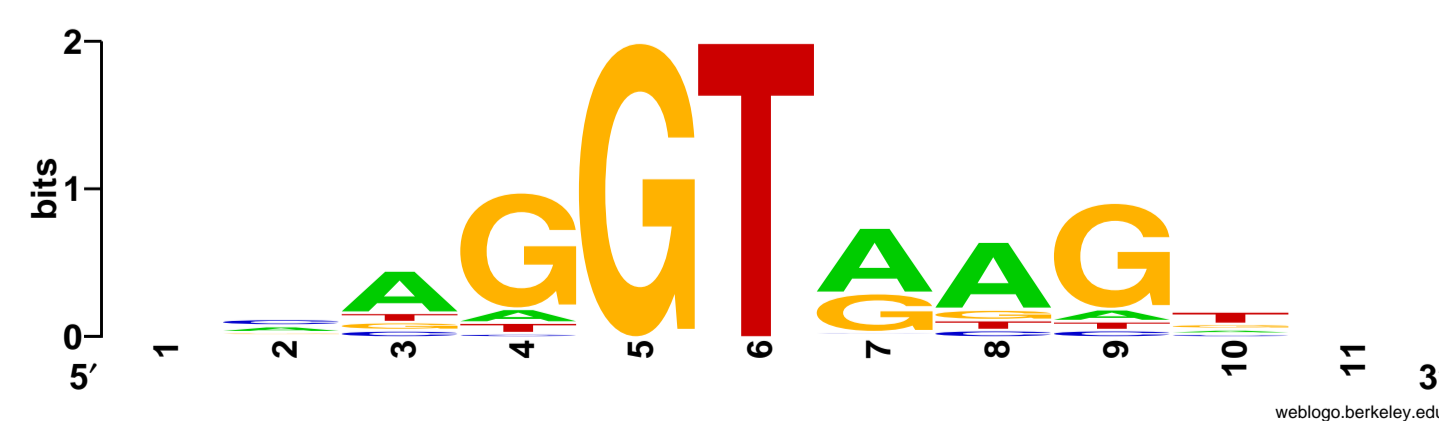


FIGURE 1: 5' SS consensus

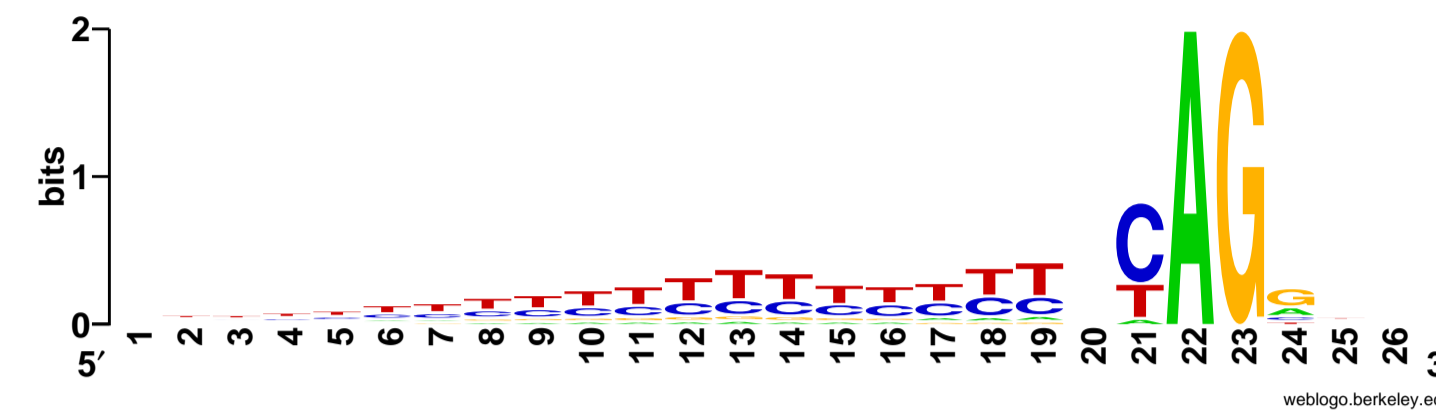


FIGURE 2: 3' SS consensus

## Existing Approaches

Different SS sensors have been developed over years, including Weight Matrix Model (WMM), Weight Array Model (WAM), windowed second-order WAM model (WWAM), Maximal Dependence Decomposition (MDD) sensor outperformed previously known sensors. It explores long-range dependencies in the donor 5' motif by iterative subdivision at each stage splitting on the most dependent position, suitably defined. Leaves of resulting bifurcation tree appear to be simple WMM models. Another approach, published in parallel and implemented in SpliceView program, explores the same idea of creating WMM motif families with clustering algorithm, which leads to dramatic performance boost compared to simple WMM and WAM.

Various sensors were built later or in parallel based on Bayesian Networks, Neural Networks and Boltzmann machine with Bahadur expansion. Also there was recent study based on Vector Support Machine (SVM). None of these methods were shown to outperform MDD for 5' SS.

New maximum entropy sensor [3] and approach based on Permuted Markov Models outperformed MDD on 5' SS.

## Proposed Methods

Our sensor design is based on 7-mer oligonucleotide counting (16,384 possible oligos) in splice and splice-like signals, with placement of 7-mer blocks within consensus similar to Maximum Entropy Sensor [3] <http://genes.mit.edu/burgelab/maxent/> as shown in Figure 3 and 4.

We used our GIGoGene [2] tool to collect extensive learning set of predicted human and mouse gene structures from where we extracted 179,079 donor and 34,258,282 donor-like signals (surrounding GT dinucleotide) plus 179,076 acceptor and 44,353,884 acceptor-like signals (surrounding AG dinucleotide).

Based on collected oligonucleotide frequencies, we can evaluate probability of a SS given an oligonucleotide

$$P(ss|oligo) = \frac{P(ss) \times P(oligo|ss)}{P(ss) \times P(oligo|ss) + P(\neg ss) \times P(oligo|\neg ss)}$$

Using the learning set, we evaluated SS interactions for signals of different strengths (on scale 1-10) and interpolated normalized signal concentration ratios  $\log_2 \left( \frac{\text{splice}}{\text{splice-like}} \right)$  to get LOD diagrams.

We incorporated found biases into our new SpliceScan tool, the curve named 'SpliceScan split-sample' in Figures 6 and 7.

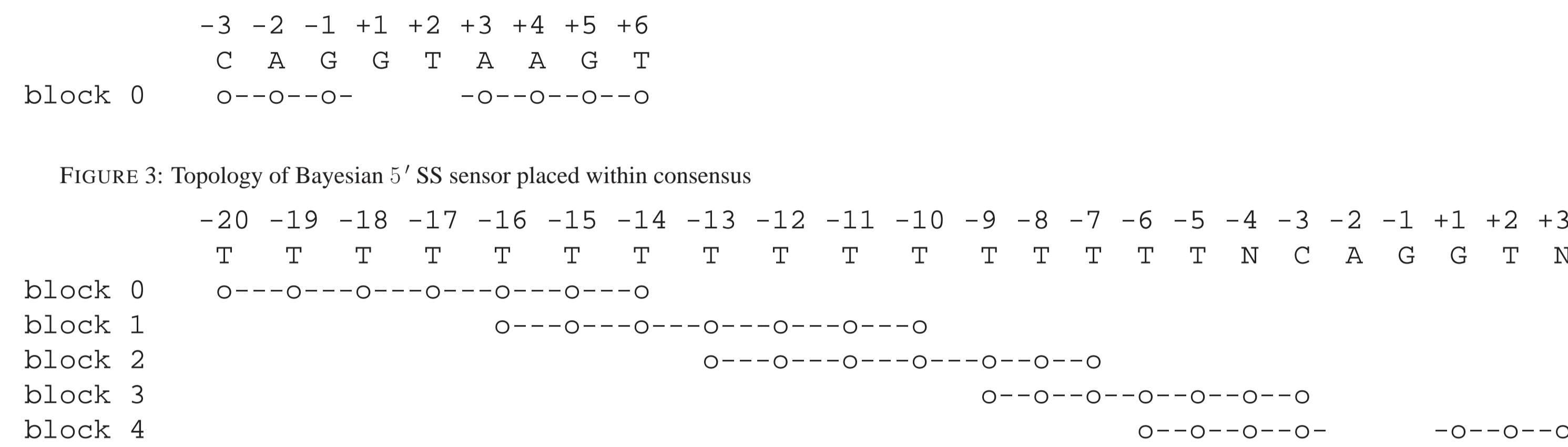


FIGURE 3: Topology of Bayesian 5' SS sensor placed within consensus



FIGURE 4: Topology of Bayesian 3' SS sensor placed within consensus

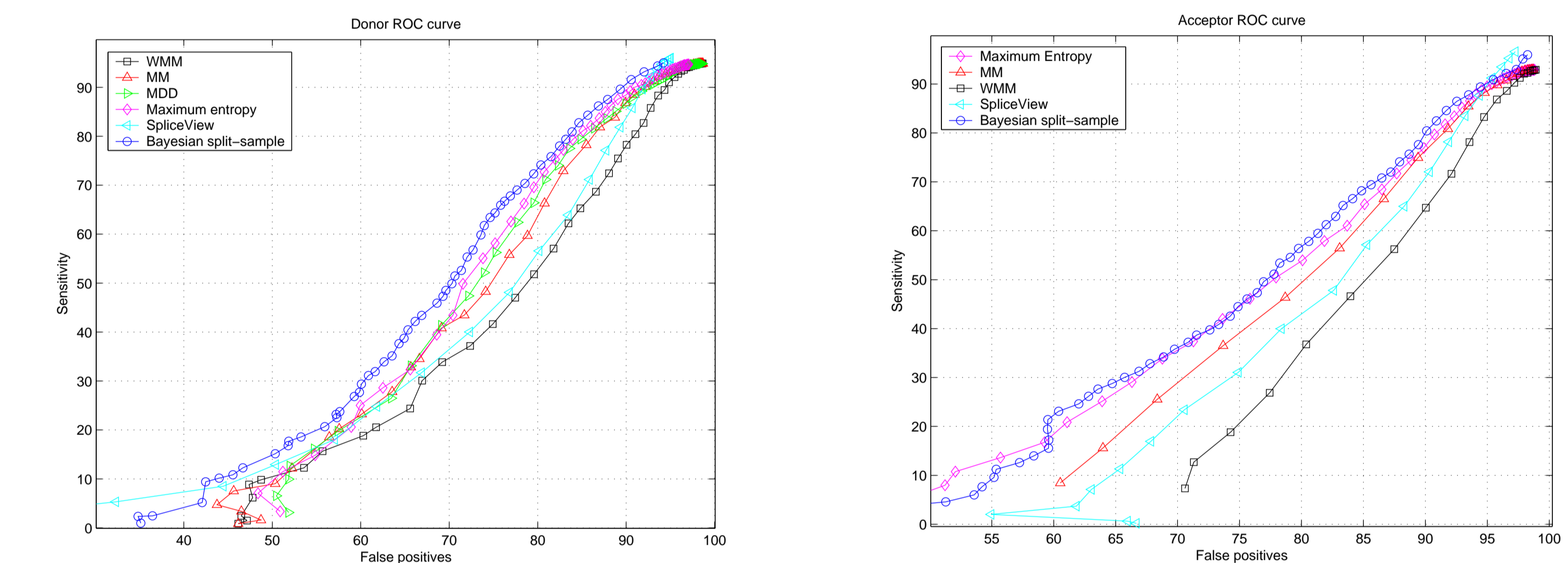


FIGURE 5: Comparison of various 5' and 3' SS sensor designs. Sets of 250 human and 183 rat genes were specifically excluded from the learning set.

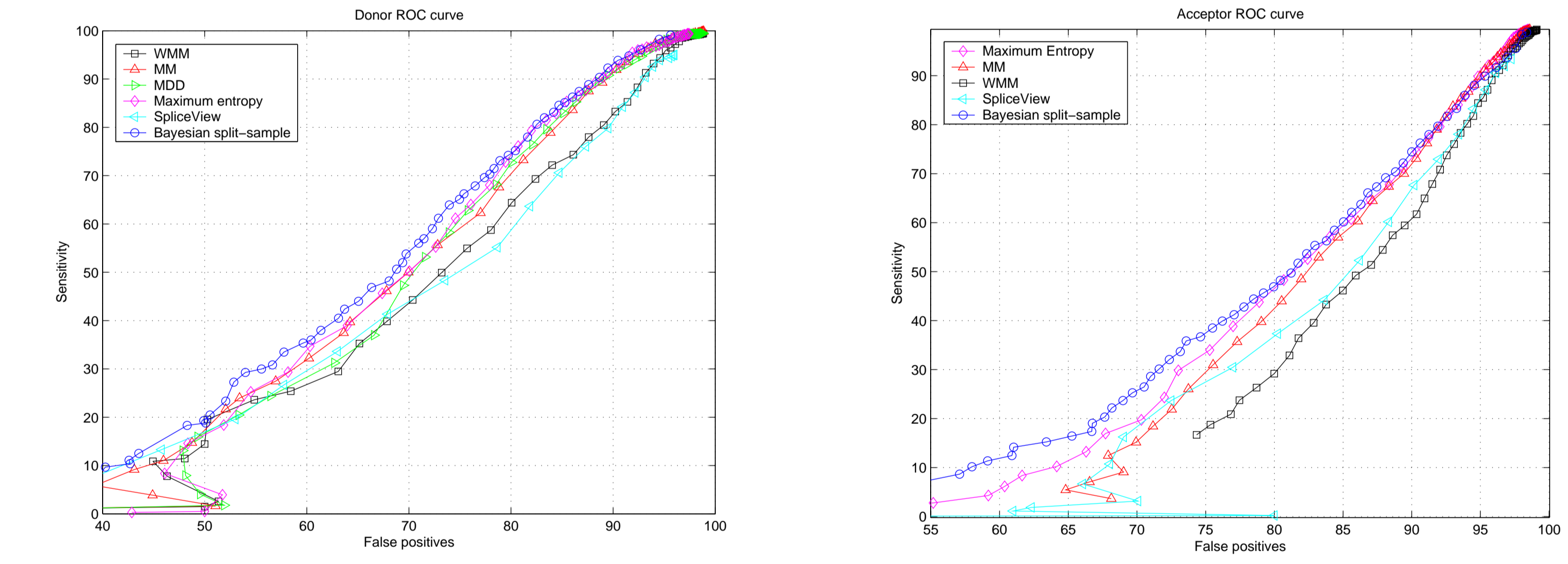


FIGURE 6: 5' ROC comparative performance for different applications

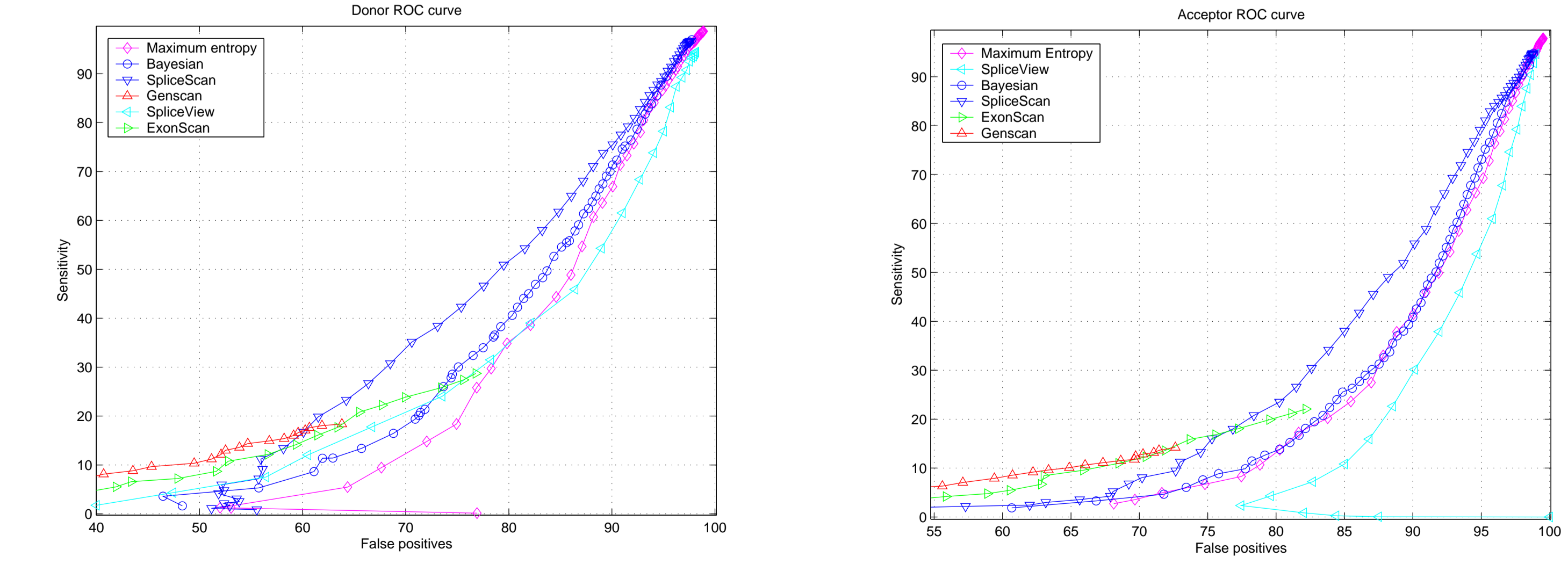


FIGURE 7: 3' ROC comparative performance for different applications

## Results and Conclusions

ROC curve is a graphical representation of the trade off between the false negative and false positive rates for every possible cut off. By tradition, the plot shows the false positive rate ( $1 - Specificity$ ) on the  $x$  axis and  $Sensitivity$  on the  $y$  axis. The closer the curve follows the left-hand and the top borders of the ROC space, the more accurate the test.

Comparative studies of prediction quality for Bayesian and Maximum Entropy sensors, shown in Figure 5, demonstrate clear advantage of our approach for the 5' and 3' SS sensor design. Figures 6 and 7 show comparative study of various tools for the set of 250 human genes, where prediction of SpliceScan is further enhanced with Intronic and Exonic Splicing Enhancers (curve named 'SpliceScan split-sample plus ISE/ESE signals').

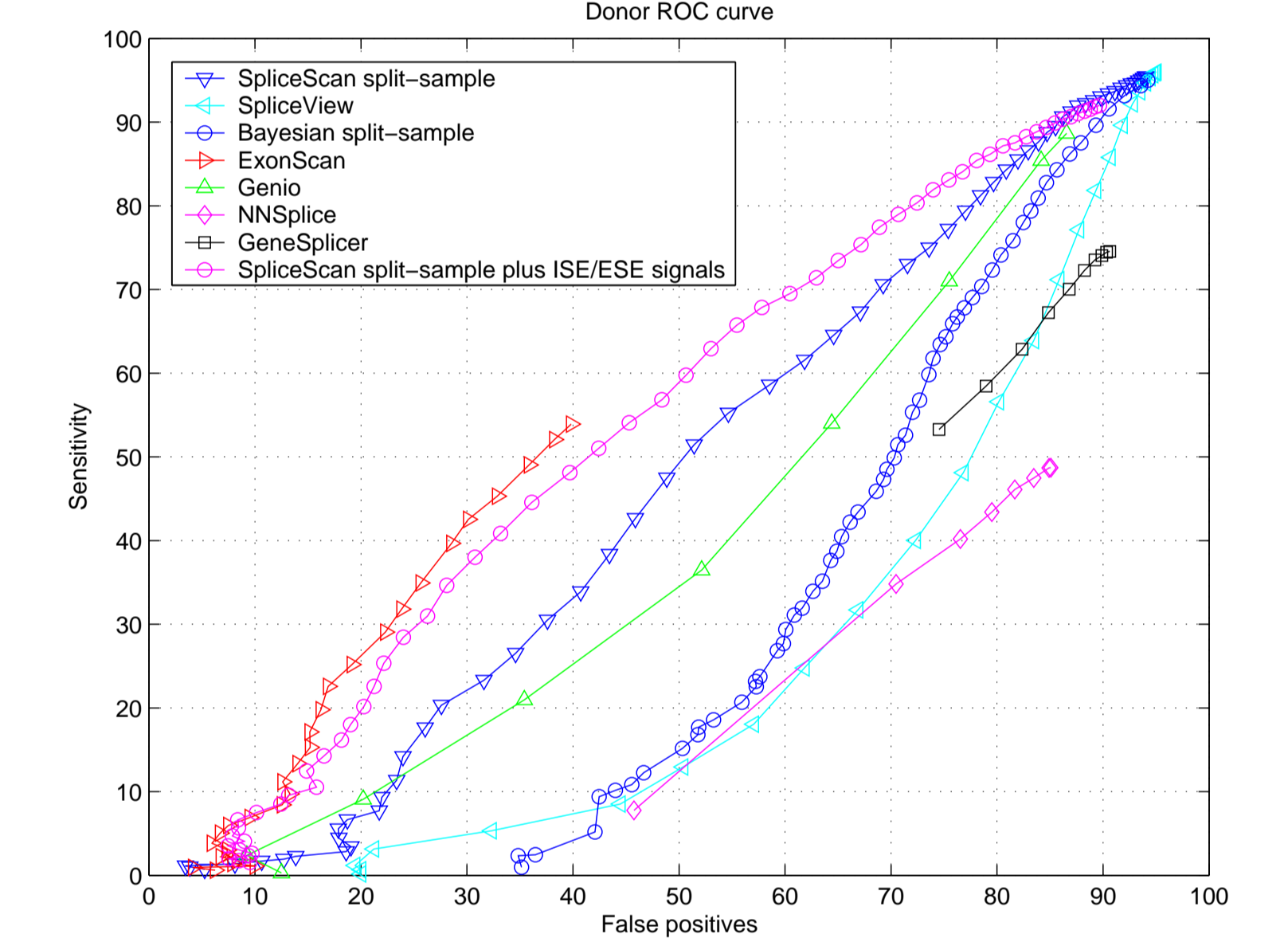


FIGURE 6: 5' ROC comparative performance for different applications

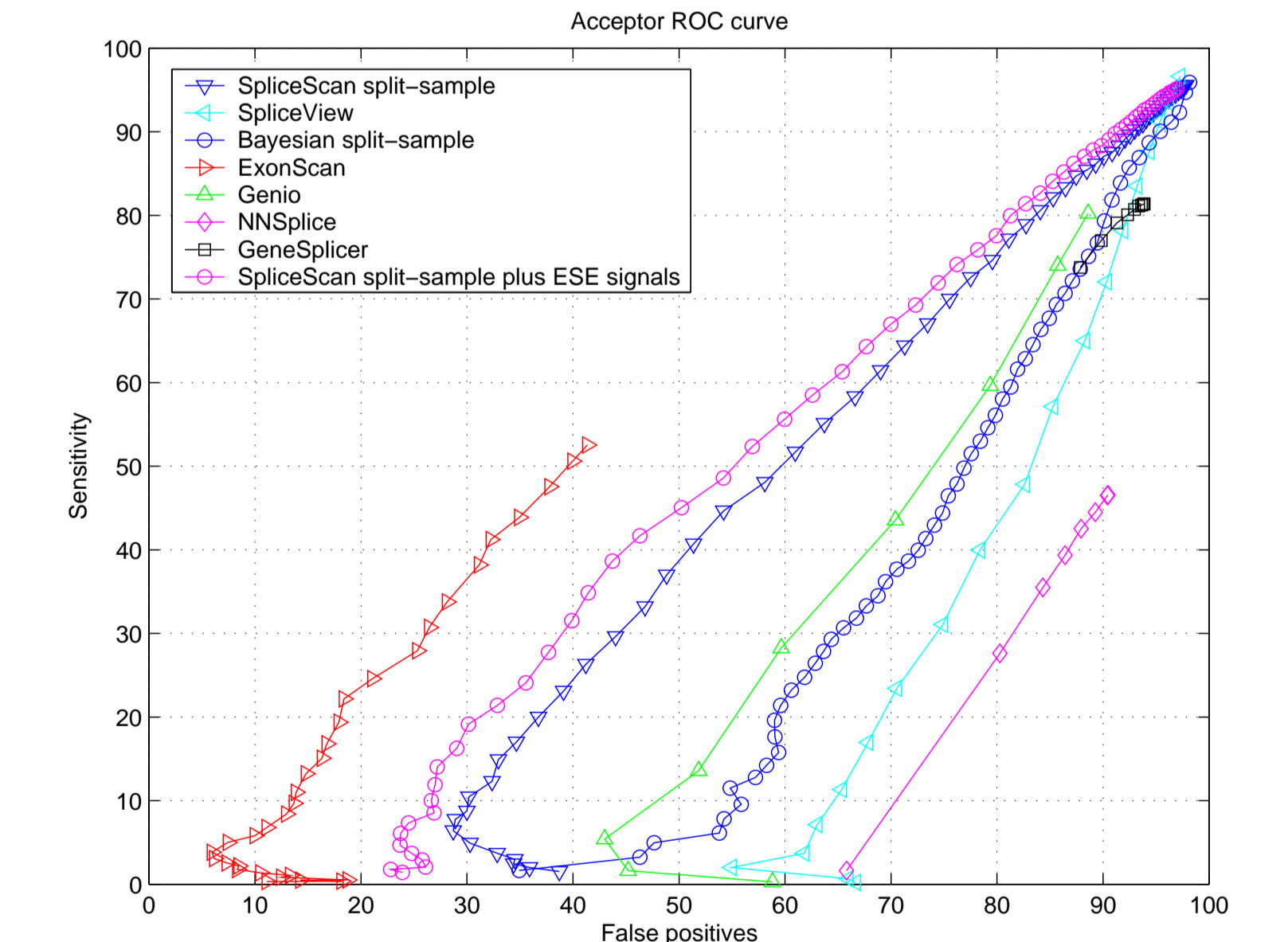


FIGURE 7: 3' ROC comparative performance for different applications

## References

- [1] A. Krogh, *Gene finding: putting the parts together*, Guide to Human Genome Computing (Martin J. Bishop, ed.), Academic Press, San Diego, CA, 2 ed., 1998, pp. 261-274.
- [2] A. Tchourbanov, D. Quest, H. Ali, M. Pauley, and R. Norgren, *A new approach for gene annotation using unambiguous sequence joining*, Proceedings of the Computational Systems Bioinformatics (CSB'03), IEEE Computer society, Aug. 2003, pp. 353-362.
- [3] G. Yeo and C.B. Burge, *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals*, Journal of Computational Biology 11 (2004), no. 2, 377-394.