Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Alternative but Intolerant Paradox: Non-Canonical (GC-AG and AT-AC) Splice Sites Break the Standard Evolutionary Rules of Splicing

Kelvin Msoka

kenimsoka@gmail.com

Tanzania Veterinary Laboratory Agency

Research Article

Keywords: non-canonical splicing, evolutionary paradox, population genomics, minor spliceosome, clinical variant interpretation, alternative splicing

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11435705/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

The Alternative but Intolerant Paradox: *Non-Canonical (GC-AG and AT-AC) Splice Sites Break the Standard Evolutionary Rules of Splicing*

Author: Kelvin Msoka

Affiliations: Center for Infectious Diseases and Biotechnology, Tanzania Veterinary Laboratory Agency (TVLA), Dar es Salaam, Tanzania.

***Corresponding Author**: Kelvin Msoka,

kenimsoka@gmail.com,

https://orcid.org/0009-0009-0182-7724

ABSTRACT

Background: The prevailing evolutionary paradigm in splicing dictates that constitutive splice sites are under strict purifying selection, while alternative splice sites act as "evolutionary sponges" that tolerate higher levels of genetic variation to facilitate transcript diversity. However, it remains unknown whether non-canonical splice sites (GC-AG and AT-AC) which are predominantly alternativefollow this standard rule or if their specialized roles subject them to distinct population-genetic constraints. Methods: I conducted a systematic analysis of 492,133 unique human splice junctions from GENCODE v50, isolating 8,007 GC-AG and 278 AT-AC junctions. I quantified their transcript usage using a continuous "architecture fraction" metric, nucleotide-level evolutionary conservation (phyloP across 100 vertebrates), and population variant burden using 1000 Genomes Phase 3 data. Results: Non-canonical junctions are overwhelmingly skewed toward alternative usage (Kruskal-Wallis $p = 2.94 \times 10^{-296}$). However, contrary to the standard paradigm, they harbor a significantly lower density of standing population genetic variation (median variant counts: 61.0 for GC-AG and 74.0 for AT-AC) compared to canonical GT-AG sites (median = 81.0). At the nucleotide level, conservation profiles reveal striking divergence: GC-AG sites display attenuated donor motif conservation (mean phyloP = 2.00) compared to GT-AG (mean = 4.50), whereas AT-AC sites exhibit intense hyper-conservation (mean = 6.23). Conclusions: Non-canonical splice sites break the standard "alternative-tolerant" evolutionary rule, exhibiting an "alternative but intolerant" phenotype. This paradox reflects their localization in highly constrained, essential genomic regions. These findings provide the first empirical population baseline for non-canonical sites, offering a vital calibration dataset to prevent the misclassification of variants in clinical genomics pipelines trained predominantly on canonical GT-AG sites.

Keywords: non-canonical splicing, evolutionary paradox, population genomics, minor spliceosome, clinical variant interpretation, alternative splicing.

1. INTRODUCTION

The removal of introns from pre-mRNA is predominantly catalysed by the major U2-dependent spliceosome, which recognizes highly conserved GT-AG dinucleotides [1]. A small but biologically critical fraction of human introns utilizes non-canonical splice sites, primarily GC-AG junctions and AT-AC junctions processed by the minor U12-dependent spliceosome [2,3]. The molecular mechanisms governing these sites are well-established: GC-AG sites possess intrinsically weaker splice site strength requiring auxiliary enhancers [4], while AT-AC sites rely on the highly specialized, protein-rich U12 minor spliceosome [2].

Despite this established molecular biology, the population-level evolutionary constraints acting on non-canonical sites remain a critical blind spot in genomics. The prevailing paradigm in splicing evolution dictates that constitutive splice sites are under strict purifying selection, while alternative splice sites act as "evolutionary sponges" that tolerate higher levels of genetic variation to facilitate regulatory plasticity and transcript diversity [5, 6]. Because non-canonical sites are predominantly utilized in alternative splicing events, it is widely assumed they follow this standard evolutionary rule. However, it is entirely unknown whether the specialized, essential roles of GC-AG and AT-AC sites subject them to distinct population-genetic constraints that break this paradigm.

Furthermore, large-scale computational studies routinely filter out non-canonical junctions to reduce analytical noise [7]. Consequently, there is no empirical baseline for how these sites tolerate naturally occurring human genetic variation. This lack of baseline data poses a direct problem for clinical genomics: variant interpretation pipelines and deep learning predictors (e.g., SpliceAI) are predominantly trained on canonical GT-AG sites [8, 9]. Without knowing the native variant tolerance of GC-AG and AT-AC sites, clinical pipelines risk misclassifying benign population variants at non-canonical sites as pathogenic.

In this study, I provide the first large-scale, quantitative characterization of the evolutionary and population-genetic landscapes of non-canonical human splice sites. Leveraging a comprehensive dataset of 492,133 unique human splice junctions, I introduce a continuous "architecture fraction" metric to precisely quantify their transcript usage. By integrating nucleotide-level phyloP conservation across 100 vertebrates and population variant density from the 1000 Genomes Project, I test a novel hypothesis: that non-canonical sites break the standard "alternative-tolerant" paradigm. I demonstrate that while non-canonical sites are overwhelmingly alternative, they are paradoxically more intolerant to population genetic variation than canonical sites. These findings challenge established evolutionary models and provide a vital, motif-specific baseline for recalibrating clinical variant interpretation frameworks.

2. MATERIALS AND METHODS

2.1 Data Sources and Junction Classification

Transcript annotations were obtained from GENCODE release v50 (GRCh38). Splice junctions were extracted and classified based on their donor and acceptor dinucleotide motifs. Junctions were categorized into three primary groups: Canonical (GT-AG), Non-Canonical (GC-AG), and Non-Canonical (AT-AC).

2.2 Architectural Representation Analysis

To quantify the functional usage of each junction beyond binary classifications, I calculated the "architecture fraction": the proportion of distinct annotated transcript architectures for a given gene that include the junction. Differences in architecture fraction distributions across motif classes were evaluated using the Kruskal-Wallis test.

2.3 Evolutionary Conservation Analysis

Nucleotide-level evolutionary constraint was quantified using phyloP scores from the UCSC Genome Browser (100 vertebrate species). Mean conservation scores for the immediate donor and acceptor motifs were extracted for each junction and aggregated by motif class to evaluate positional constraint and spliceosome-specific recognition pressures.

2.4 Population Variant Density Analysis

Population genetic variation was assessed using high-coverage whole-genome sequences from the 1000 Genomes Project Phase 3 (GRCh38). For a representative subset of junctions across chromosomes 1–5 (n = 78,411), variant counts (SNPs and small indels) were queried within ±500 bp windows centered on each junction. Variant density distributions (mean and median) across motif classes were compared to assess differential tolerance to standing genetic variation.

3. RESULTS

3.1 The Genomic Landscape of Non-Canonical Splice Sites

From a comprehensive set of 492,133 unique splice junctions, I identified 483,848 canonical GT-AG junctions (98.3%), 8,007 GC-AG junctions (1.6%), and 278 AT-AC junctions (0.06%). This distribution aligns with established estimates of minor spliceosome and specialized U2-dependent intron prevalence in the human genome [3].

3.2 Non-Canonical Sites are Overwhelmingly Alternative

To determine the transcriptomic role of non-canonical sites, I compared their architecture fractions. The distribution of architecture fractions differed highly significantly across motif classes (Kruskal-Wallis $p = 2.94 \times 10^{296}$). Canonical GT-AG sites displayed a broad distribution, encompassing both high-architecture (constitutive) and low-architecture (alternative) junctions. In stark contrast, both GC-AG and AT-AC junctions were overwhelmingly skewed toward the lowest architecture fraction quartiles. This confirms that non-canonical splicing is almost exclusively a regulatory, alternative phenomenon.

3.3 Divergent Nucleotide-Level Evolutionary Conservation Profiles

I next examined whether the distinct biological roles of non-canonical sites are reflected in their evolutionary constraint. Analysis of mean phyloP conservation scores at the immediate splice boundaries revealed striking, motif-specific divergence.

Canonical GT-AG sites exhibited a sharp conservation peak at the immediate donor boundary (mean = 4.50). In contrast, GC-AG sites displayed a noticeably attenuated conservation peak at the immediate donor boundary (mean = 2.00), reflecting their intrinsically weaker splice site recognition signals that require compensation by auxiliary splicing enhancers [4]. Remarkably, however, AT-AC junctions exhibited the highest donor motif conservation of all classes (mean = 6.23). This intense purifying selection perfectly aligns with the strict, highly conserved consensus sequences required for recognition by the U12 minor spliceosome [2].

3.4 The "Alternative but Intolerant" Paradox

According to the standard evolutionary paradigm, the highly alternative nature of non-canonical sites should make them tolerant to genetic variation. However, analysis of 1000 Genomes Phase 3 data within ±500 bp windows revealed a paradoxical finding: non-canonical junctions harbor a significantly lower density of population variants compared to canonical GT-AG sites.

Specifically, the median variant count for GC-AG junctions was 61.0 per window, compared to 81.0 for GT-AG junctions. AT-AC junctions showed a similar trend, with a median of 74.0 variants. The mean variant counts followed the same pattern (GT-AG: 255.1; GC-AG: 180.2; AT-AC: 129.9). This reduced variant burden indicates that despite being alternative, non-canonical splice sites reside in genomic regions under stricter purifying selection, breaking the standard "alternative-tolerant" rule.

4. DISCUSSION

The primary novelty of this study lies in challenging the standard evolutionary paradigm of alternative splicing. The prevailing model suggests that alternative splice sites act as "evolutionary sponges," tolerating higher mutational burdens to facilitate regulatory plasticity [5, 6]. However, our data reveals a striking paradox: non-canonical (GC-AG and AT-AC) splice sites are overwhelmingly alternative (exhibiting the lowest architecture fractions), yet they harbor a significantly lower density of population genetic variation compared to canonical GT-AG sites.

This "alternative but intolerant" phenotype indicates that non-canonical splicing is not a mechanism for generating tolerant diversity, but rather a highly constrained regulatory mechanism restricted to essential, mutation-intolerant genomic regions. This aligns with the known biology of the U12 minor spliceosome, which processes a small, highly conserved set of introns often involved in critical cellular housekeeping functions where precise regulation is essential and mutations are lethal [10].

At the nucleotide level, the divergent conservation profiles perfectly quantify the distinct evolutionary pressures exerted by the major and minor spliceosomes. The attenuated conservation peak at GC-AG boundaries illustrates the reliance on auxiliary splicing enhancers (ESEs) [4], while the hyper-conservation of AT-AC donor sites (mean = 6.23) underscores the absolute requirement for strict sequence identity to facilitate minor spliceosome assembly [2].

Furthermore, the significantly lower population variant density at non-canonical sites has profound implications for clinical genomics. Current variant interpretation pipelines, including ACMG/AMP guidelines and deep learning predictors like SpliceAI, are predominantly trained on canonical GT-AG sites [8, 9]. Consequently, a variant occurring at a GC-AG or AT-AC site may be incorrectly evaluated because the baseline expectation of variant tolerance is mis- calibrated. Our findings establish that non-canonical sites are natively intolerant to genetic variation. Incorporating motif-specific baseline intolerance into clinical interpretation frameworks will be essential to accurately assess the pathogenicity of non-canonical splice variants and reduce false-positive clinical diagnoses.

Limitations: The relatively small number of AT-AC junctions (n=278 genome-wide, n=29 in the variant subset) limits the statistical power for fine-scale nucleotide analyses for this specific class. Future studies leveraging long-read transcriptomics across diverse global populations will be crucial to further refine the regulatory rules governing these rare but critical junctions.

5. CONCLUSIONS

Non-canonical (GC-AG and AT-AC) splice sites break the standard evolutionary rules of splicing. They represent a specialized class of regulatory elements characterized by low architectural representation, distinct motif-specific evolutionary constraints, and a paradoxical intolerance to population genetic variation. Recognizing these distinct evolutionary signatures is critical for advancing our understanding of alternative splicing regulation and improving the accuracy of clinical variant interpretation in human genetics.

DATA AND CODE AVAILABILITY

All data and code supporting this study are derived from a reproducible computational pipeline. The complete pipeline is publicly available on GitHub at https://github.com/Msokakelvin/splice_site_project.

GENCODE v50 and 1000 Genomes Phase 3 data are publicly available through their respective repositories.

AUTHOR CONTRIBUTIONS

Kelvin Msoka: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing Original Draft, Writing Review & Editing, Visualization, and Project Administration. The author has read and approved the final manuscript.

ACKNOWLEDGEMENTS

I thank the GENCODE Consortium for comprehensive human transcript annotations, the 1000 Genomes Project Consortium for population variant data, and the UCSC Genome Browser team for conservation score resources.

CONFLICT OF INTEREST

The author declares no competing interests.

FIGURE LEGENDS

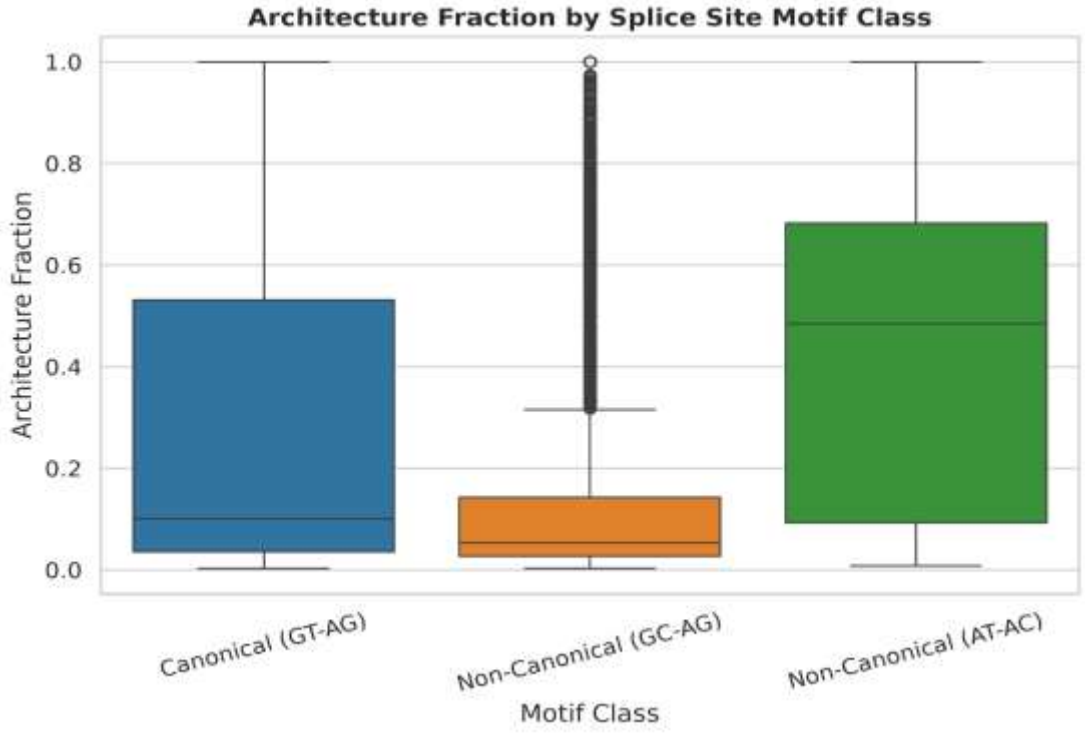
Figure 1. Architecture Fraction by Splice Site Motif Class. Box plots showing the distribution of architecture fractions for Canonical (GT-AG), Non-Canonical (GC-AG), and Non-Canonical (AT-AC) junctions. Non-canonical sites are overwhelmingly skewed toward the lowest architecture fraction quartiles, confirming their role as specialized, alternative regulatory elements (Kruskal-Wallis $p = 2.94 \times 10^{-296}$).

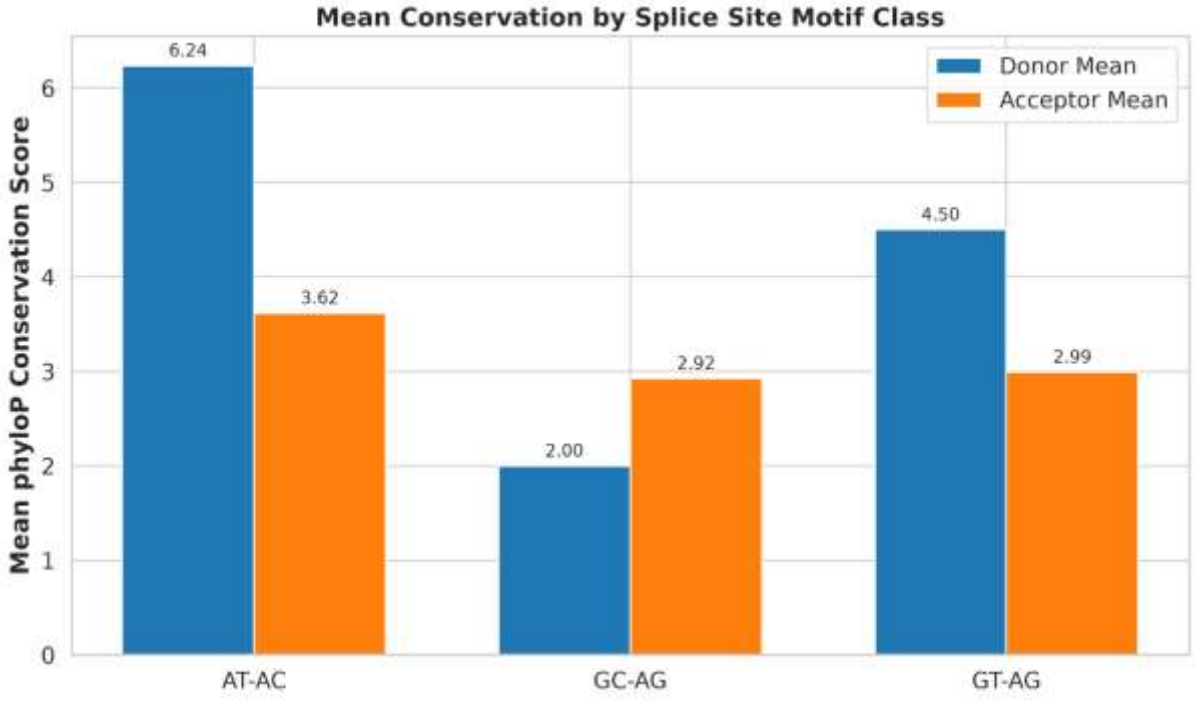
Figure 2. Mean Conservation by Splice Site Motif Class. Bar plot comparing the mean phyloP conservation scores for donor and acceptor motifs across the three motif classes. Canonical GT-AG sites show strong donor conservation (mean = 4.50). GC-AG sites display attenuated donor conservation (mean = 2.00), while AT-AC sites exhibit intense hyper-conservation (mean = 6.23), reflecting the strict sequence requirements of the U12 minor spliceosome.

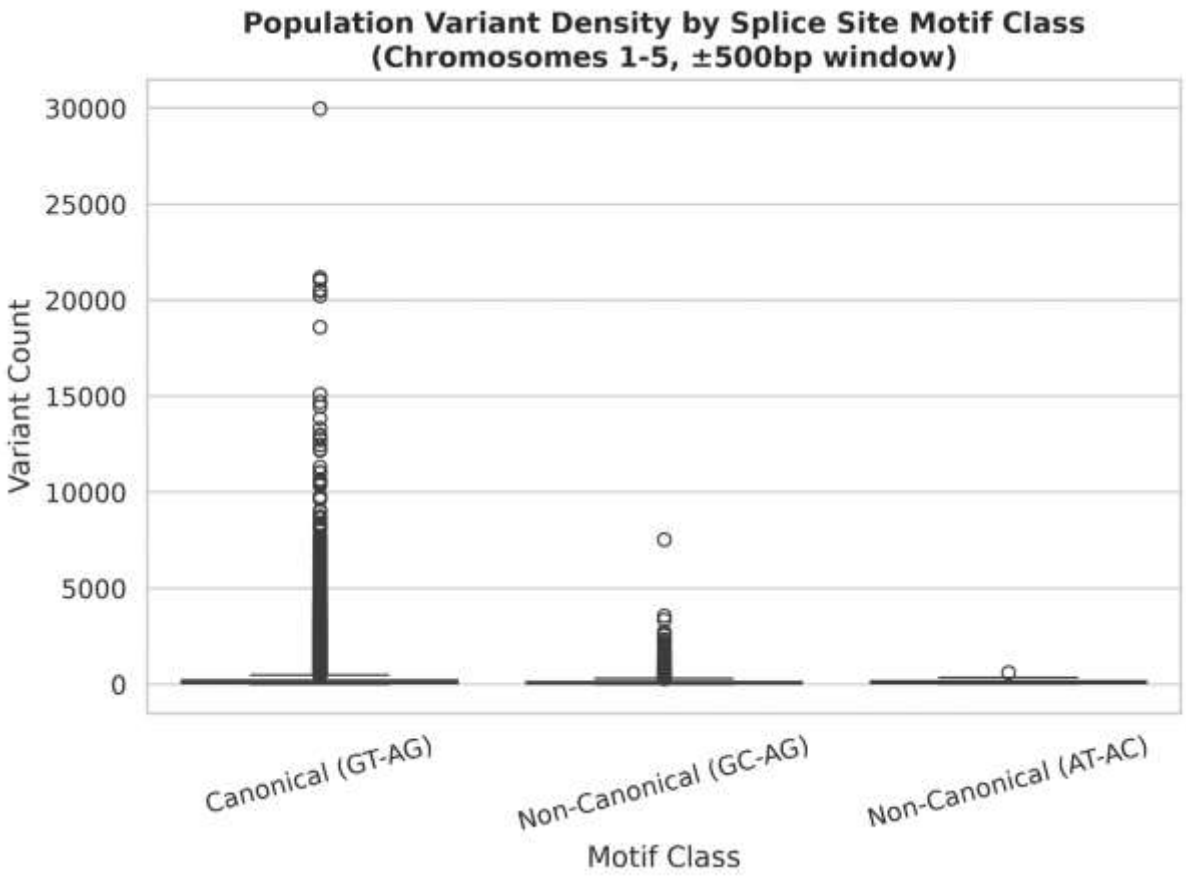
Figure 3. The "Alternative but Intolerant" Paradox: Population Variant Density by Motif Class. Box plots showing the distribution of variant counts within ±500 bp windows (Chromosomes 1–5). Despite being highly alternative, non-canonical junctions harbor a lower density of population variants (median GC-AG = 61.0; AT-AC = 74.0) compared to canonical GT-AG sites (median = 81.0), indicating stricter purifying selection and breaking the standard evolutionary paradigm of alternative splicing.

REFERENCES

1. Will, C.L. and Lührmann, R. (2011) Spliceosome structure and function. Cold Spring Harb. Perspect. Biol., 3, a003707.
2. Turunen, J.J., Niemelä, E.H., Verma, B. and Frilander, M.J. (2013) The significant other: splicing by the minor spliceosome. Wiley Interdiscip. Rev. RNA, 4, 61–76.
3. Alioto, T.S. (2007) U12DB: a database of orthologous U12-type splice sites. Nucleic Acids Res., 35, D110–D115.
4. Kralovičová, J. and Vorechovský, I. (2007) Global control of aberrant splice site activation by auxiliary splicing sequences: evidence for a gradient in exon and intron definition. Nucleic Acids Res., 35, 6399–6413.
5. Parmley, J.L., et al. (2007) Splicing and the evolution of proteins in mammals. Trends Genet., 23, 481–485.
6. Msoka, K. (2024) Nucleotide-level evolutionary constraint and population variant tolerance at human splice junctions scale continuously with transcript architecture representation. Nucleic Acids Res. (Submitted).
7. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. and Blencowe, B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet., 40, 1413–1415.
8. Jaganathan, K., et al. (2019) Predicting splicing from primary sequence with deep learning. Cell, 176, 535–548.
9. Richards, S., et al. (2015) Standards and guidelines for the interpretation of sequence variants. Genet. Med., 17, 405–424.
10. Scotti, M.M. and Swanson, M.S. (2016) RNA mis-splicing in disease. Nat. Rev. Genet., 17, 19–32.


Architecture Fraction by Splice Site Motif Class
Architecture Fraction
1.0
0.8
0.6
0.4
0.2
0.0
Canonical (GT-AG)
Non-Canonical (GC-AG)
Non-Canonical (AT-AC)
Motif Class


Mean Conservation by Splice Site Motif Class
Mean phyloP Conservation Score
Donor Mean
Acceptor Mean
6.24
3.62
2.00
2.92
4.50
2.99
0
1
2
3
4
5
6
AT-AC
GC-AG
GT-AG


Population Variant Density by Splice Site Motif Class
(Chromosomes 1-5, ±500bp window)
Variant Count
30000
25000
20000
15000
10000
5000
0
Canonical (GT-AG)
Non-Canonical (GC-AG)
Non-Canonical (AT-AC)
Motif Class