Using Human Genetic Variation to Predict Functional Elements in Non-Coding Genomic Regions

by

David Lomelin

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

Committee in Charge

Deposited in the Library, University of California, San Francisco

Date University Librarian

Degree Conferred:

Using Human Genetic Variation to Predict Functional Elements in Non-Coding Genomic Regions

David Lomelin

Abstract

The annotation of the human genome has been a daunting task requiring the creation of innovative methods to characterize its diverse elements. Given that previous studies have successfully used human polymorphism data to characterize functional elements within coding regions, the objective of this thesis is to use human polymorphism data to improve the identification of functional elements in both coding and non-coding regions. This study relies on using the combination of genetic variation from ethnically diverse human populations and several bioinformatics approaches to discriminate and identify several elements of functional importance within genomic regions.

Human polymorphism data within genes was acquired from three different publicly available datasets. We then demonstrated that positions within introns that correspond to known functional elements involved in pre-mRNA splicing, including the branch site, splice sites, and polypyrimidine tract showed reduced levels of genetic variation. These precise sites of reduced polymorphism levels also coincide with the positions known for base pairing and interacting with their corresponding ligand. Furthermore, we observed regions of reduced genetic variation that were candidates for distance dependent localization sites of functional elements. Using several computational approaches, we provided additional evidence that suggests these regions correspond to intronic splicing enhancers in both the 5’ and 3’ splice site regions.

We conclude that studies of genetic variation can successfully discriminate and identify functional elements in non-coding regions. Although current polymorphism data is only available for small gene subsets, as more non-coding sequence data becomes available, the methods employed here can be utilized to identify additional functional elements in the human genome and provide possible explanations for phenotypic associations.

Table of Contents

List of Tables

List of Figures

List of Abreviations

A	adenine
bp	basepair
C	cytosine
DNA	deoxyribonucleic acid
G	guanine
GWA	genome wide association
hnRNA	heteronuclear ribonucleic acid
HMM	Hidden Markov model
ISE	intronic splicing enhancer
mRNA	messenger ribonucleic acid
PMT	Pharmacogenetics of Membrane Transporters
PPT	polypyrimidine tract
PSSM	position specific scoring matrix
R	purine
RNA	ribonucleic acid
SNP	single nucleotide polymorphism
SS	splice site
snRNA	small nuclear ribonucleic acid
snRNP	small nuclear ribonucleoprotein particle
T	thymine
U	uracil
Y	pyrimidine

1 Introduction

With recent advances in sequencing technology, the genomes of 5,334 species[1], including humans[2], have been sequenced. While this provides a wealth of data for the scientific community, these sequences must be annotated to provide usable information for research purposes. Even with current technology, characterizing the functional elements within genomes across different species remains a difficult task. Given our knowledge that the gene is the genome's basic functional unit, a lot of research and effort has been placed on developing computational methods to facilitate the recognition of these coding regions. While these coding regions are relatively straight forward to identify, we lack an understanding as well as a tool set to identify similarly important non-coding regions. Therefore, the purpose of this thesis is to improve upon and develop techniques for the recognition of functional non-coding regions through the use of computational methods.

Experimental methods used for gene identification rely on the sequencing of mRNAs and ESTs, which provide the sequences of concatenated exons from genes. These sequences are then mapped and aligned onto genomic sequence through computational methods. While these powerful techniques can characterize gene positions and structures by resolving intron-exon boundaries, the alignment of known ESTs to a well annotated genome, such as Arabidopsis, was shown to overlap of up to 70% of its known genes (personal work, unpublished). Therefore, computational methods must be implemented to complement experimental genome annotation methods.

There are a number of different computational methods designed for gene prediction, which differ by the type of input data they require and the type of algorithm that they implement. The input data can be in the form of sequence signals such as splicing motifs, content statistics such as codon and nucleotide usage, and similarity to known genes in public databases[3]. While there are a diverse number of algorithms, most gene finding programs are focused on classifying protein-coding regions as opposed to noncoding RNA or regulatory regions[4]. The best current methods are known to have high sensitivity and specificity measures above 90% when predicting if a given nucleotide is part of the coding part of a gene; however, when more stringent criterion are applied, such as predicting exon boundaries or complete gene structures, sensitivity and specificity measures generally drop below 40%[5]. This drop in performance suggests a lack of a complete understanding of the characteristics that define the exact gene boundaries and our ability to properly model genes.

Other computational methods include comparative genome algorithms where genomes from closely related species, such as human and mouse, are compared and characterized for synteny[6]. When these comparisons are performed within recently diverged species, several conserved regions are often highlighted which are indicative of functional coding and noncoding elements. Such alignments have successfully recognized protein-coding genes, non-coding RNA genes, regulatory regions, and DNA replication sites[7,8]; however, some studies using the same multi-species methods have identified conserved non-genic sequences of unknown function that are assumed to be preserved because of their functional importance[9,10]. To address the importance of these potentially functional non-coding segments, an experiment was performed where four such ultraconserved elements were removed from the mouse genome[11]. Surprisingly, this resulted in viable mice that had no observable phenotypic effects. This is just one potential downfall of using comparative genomics, in addition to the fact that results from this technique are highly dependent on the similarity between the organisms being compared. Additionally, these methods have the potential to overlook motifs specific to one of the species being compared.

Given the lack of a complete understanding about the structural boundaries between functional and non functional genomic regions, the shortage of non protein-coding functional annotation programs, and the poor results from current gene prediction algorithms, this thesis will present the implementation of a generalized method for predicting functional elements within noncoding regions. The main idea behind this thesis comes from studies done in protein coding regions in the Giacomini lab where sequence variants in positions that led to nonsynonymous substitutions were found to have lower allele frequencies compared to other sequence changes[12]. Follow up work by the same group measured the in vitro activity of polymorphic transmembrane transporters and showed that the transporters with high frequency variants retained function while those with low frequency variants more often lost activity or displayed reduced function[13]. This suggests that alleles that are functionally deleterious will be selected against and thus underrepresented among high frequency variants and over-represented among low frequency variants. Sites with decreased polymorphic levels, that is, having a low variant frequency, have been suggested to be under purifying selection and therefore likely indicate functionally important regions within the genome. Given that these previous studies have successfully used human polymorphism data to characterize functional elements within human coding regions, the objective of this thesis is to show that biologically active sites within non-coding regions show the same reduced genetic variation characteristics that are seen within coding regions.

Chapter 2 presents the initial study on the genetic variation levels found in introns - these noncoding regions are known to contain functionally important motifs critical for gene expression and RNA splicing. This study shows that the functional motifs within introns that are important for cell function have reduced levels of genetic variation. In addition, three regions within introns also have reduced polymorphism levels but do not match the common functional intronic elements. Evidence from the literature suggests that one region downstream of the 5' splice site corresponds to a distance dependent localization site for intronic splicing enhancers. Based on knowledge about the location for branch site motifs, we hypothesize the remaining two regions in the 3' splice site are two preferential locations for branch sites. This study concludes that functional elements within coding and noncoding regions show similar signatures of reduced polymorphism levels.

Chapter 3 investigates the properties and nature of the three regions of reduced genetic variation found previously in chapter 2. As described briefly, we hypothesized that the two regions upstream of the 3’ splice site were preferential locations for branch sites. Given the branch sites' variable positions and the degeneracy of their motif, several computational methods were implemented to predict their location. Through the measurement of the polymorphism levels of the predicted branch sites' positions, we concluded that two sites are critical for splicing, which agrees with known evidence in the literature. Additionally, the distribution of predicted branch site locations overlaps only one of the two predicted functional regions. To further characterize the properties of these two regions, a genetic variation based motif finder was developed and indicated that one region is the preferred location of branch sites while the other region is a distance dependent localization site for intronic spicing enhancers. The same motif finder was applied to the predicted functional region in the 5’ splice site region and reinforced the original prediction that this region is a distance dependent localization site for intronic splicing enhancers. This study confirms that the original predictions made in chapter two are functional non-coding elements and further exemplifies how genetic variation can be a powerful tool for characterizing novel motifs.

Given the successful use of genetic variation to identify functional motifs in non-coding regions, chapter 4 presents the study done to improve the characterization of functional elements in coding regions. In this section, we show how reduced levels of genetic variation in synonymous positions or increased levels in nonsynonymous sites can be indicators of functionally important sites. We conclude that the measurement of genetic variation in non-coding and coding regions can be a useful tool to aid in the characterization of functional regions across the human genome.

2 Measuring Genetic Variation in Non-Coding Regions

At first glance, the measurement of the genetic variation within functional non-coding regions of the human genome might appear to be rudimentary; however, there are limitations that prevent a full scale analysis of the entire genome. First, the measurement of intra-human variation requires sequenced genomes from at least 50 individuals due to the fact that sequencing more individuals yield better coverage of existing SNPs. For this reason, along with the current monetary limitations required for full scale sequencing, only the genes of interest from different institutions have been sequenced for more than one individual thereby limiting the available data for this type of analysis.

While the broader goal is to analyze intergenic non-coding regions to improve sequence annotation methods, having only gene data available restricts the non-coding regions we can characterize. Therefore, this analysis will be limited to introns, which contain non-coding motifs that are required for recognition and binding to multiple spliceosome elements during intron splicing. This chapter will therefore present information on the sequence and position of the functional elements within the intron and the basis of their interaction with the spliceosome. There will also be descriptions of the different databases that were used for analysis, the population genetic parameters used for quantifying the genetic variation, the results of the analysis as well as the results for different ethnic groups within each population, and finally the method for classifying the statistical significance of the results.

2.1 Abstract

Non-coding DNA, particularly intronic DNA, harbors important functional elements that affect gene expression and RNA splicing. Yet, it is unclear which specific non-coding sites are essential for gene function and regulation. To identify functional elements in non-coding DNA, we characterized genetic variation within introns using ethnically diverse human polymorphism data from three public databases, PMT, NIEHS, and Seattle SNPs. We demonstrate that positions within introns corresponding to known functional elements involved in pre-mRNA splicing, including the branch site, splice sites, and polypyrimidine tract show reduced levels of genetic variation. Additionally, we observed regions of reduced genetic variation that are candidates for distance dependent localization sites of functional elements, possibly intronic splicing enhancers (ISEs). While ISEs have previously been characterized in a select few genes, our findings suggest that they are found within most introns of genes. We conclude that studies of genetic variation can successfully discriminate and identify functional elements in non-coding regions. As more non-coding sequence data becomes available, the methods employed here can be utilized to identify additional functional elements in the human genome and provide possible explanations for phenotypic associations.

2.2 Introduction

Genome-wide association studies have begun to identify large numbers of genetic variants that influence the risk of human diseases and variability in human traits. A striking feature of the newly associated variants is that the top signals often occur at DNA sites that do not encode amino acids. Because the function of non-coding DNA is less well understood than that of coding DNA, researches are left to speculate about the functional effect of these variants. Methods that elucidate the function of non-coding DNA can complement the knowledge gained from association studies and in so doing lead to a more complete understanding of gene function and disease etiology. Here, we examine the distribution of genetic variation that exists within the human species to identify functional elements in the human genome.

The most common differences in DNA sequence between individuals are single nucleotide polymorphisms (SNPs), that is, changes to a single DNA basepair. A polymorphism is defined as a DNA variant whose minor (less common) allele has a frequency of at least 1%. SNPs with a minor allele frequency of 1% or greater occur on average once every 250 to 2000 nucleotides[14,15]. Sites with low genetic diversity have been suggested to be under purifying selection and therefore indicate functionally important regions within the genome. Consequently, calculation of nucleotide diversity, which provides a measure of genetic variation, is commonly employed to recognize functional sites and to characterize genetic variation[12,13,16,17,18].

To this end, surveys of DNA variation in humans have been undertaken to better understand the characteristics of functional sites in the genome. To date, most of these surveys have focused on genes. A survey of variation in 106 genes associated with cardiovascular disease, endocrinology, and neuropsychiatry, and another of 75 genes involved in blood pressure homeostasis and hypertension, showed that there was reduced variation at nonsynonymous (change in amino acid) sites within coding regions, particularly when the changes led to non-conservative (change in amino acid whose biochemical properties differ from the native amino acid) mutations[17,18]. Additional work on 24 human membrane transporter genes showed that sequence variants in positions that lead to nonsynonymous substitutions have lower allele frequencies when compared with other sequence changes[12]. Follow up work by the same group measured the *in vitro* activity of polymorphic transmembrane transporters, revealing that the transporters with high frequency variants retained function while those with low frequency variants more often lost activity or displayed reduced function[13]. In general, alleles that are functionally deleterious will be selected against and thus underrepresented among high frequency variants, and over-represented among low frequency variants.

Supported by experiments like those described above, many have emphasized studying coding as opposed to non-coding variants for phenotypic effects since these changes can have a direct effect on the protein sequence, and therefore are more likely to alter its function[19]. There are, however, also clear examples of mutations in non-coding regions, including those within introns, being responsible for diseases. Experiments in yeast have shown that mutations at the 3’ and 5’ splice sites as well as the conserved adenosine residue within the branch site, a motif that is required for intron splicing, cause aberrant splicing or prevent it altogether[20,21]. A more recent experiment in humans showed that a G→C mutation in the 3’ splice site can result in one of two outcomes: either the use of a new AG site 15bp downstream from the original splice site, that causes a frameshift and a premature stop, or skipping the exon following the 3’ splice site[22]. Figure 2-1 displays intron structure and the relative location of its known sequence elements.

Functionally important intronic variants such as those described above have also been shown to cause disease. For example, familial hypercholesterolemia has been shown to result from the removal of a branch site from intron 9 of the low density lipoprotein receptor gene[23]. Other mutations within the branch site have been shown to cause other human disorders, such as fish-eye disease, which is caused by a T→C mutation within the branch site of a lecithin-cholesterol acetyltransferase gene preventing intron removal[24]. Additionally, many genome wide association (GWA) studies have now found associations between intronic variants and diseases such as breast cancer and diabetes[25,26,27].

An important problem in GWA studies is that even after a locus is implicated in causing a disease, using linkage disequilibrium to associate a specific non-coding variant with the disease can be inconclusive since these non-coding variants typically don’t offer any information about the functional effect of the DNA change, whereas changes in coding DNA have clearer consequences[28]. Given that previous studies have successfully used human polymorphism data to characterize functional elements within human coding regions, the present study demonstrates that biologically active sites within non-coding regions, specifically introns, show the same reduced genetic variation characteristics that are seen within coding regions. In addition, we use human polymorphism data to identify novel location-specific intronic sites that suggest the presence of functional elements within non-coding DNA, which may represent intronic splicing enhancers.

2.2.1 Intron Structure

A typical eukaryotic gene is composed of several short coding sequences (exons) interspersed with longer non-coding regions (introns) (Figure 2-1). After the cell transcribes a heteronuclear RNA (hnRNA or pre-mRNA) from a gene, the intronic regions must be removed by the cell's splicing machinery before its final form (mRNA) can be used for translation – a process commonly known as RNA splicing. The complex responsible for this task is the spliceosome, which is composed of five small nuclear RNAs (snRNAs), U1, U2, U4, U5, and U6, and more than 60 polypeptides that must precisely recognize the 5' and 3' intron edges (splice sites) to properly excise the intron from the mRNA[29]. The snRNAs form a complex with proteins known as small nuclear ribonucleoprotein particles (snRNPs). Any mistakes in this process lead to aberrantly spliced mRNAs which are mistranslated.

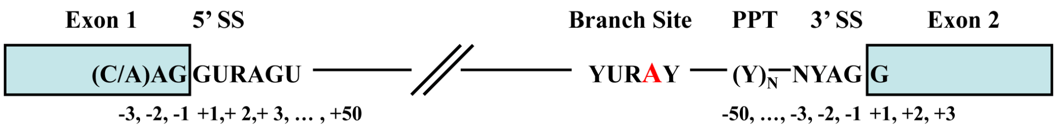

Figure 2-1 Known conserved motifs within the intron. Numbering system is relative to the intron/exon boundary of each splice site and is not to scale with the consensus sequences above. SS: Splice Site. PPT: Polypyrimidine Tract.

There are several elements within introns that associate with a number of factors from the spliceosome. Analysis of a large number of pre-mRNAs has shown that there are consensus sequences at the 5' and 3' splice sites that are highly conserved and promote spliceosome assembly: (1) the 5' splice site, characterized by the conserved sequence 5'-GURAGU-3' where the first 2 residues are particularly conserved (> 99%) across eukaryotic introns (and R denotes purine); (2) the 3' splice site, with the conserved consensus sequence 5'-NYAG-3' (where Y denotes pyrimidine and N denotes all nucleotides); and (3) a region upstream of the 3' splice site known as the polypyrimidine tract (PPT), which is a stretch of ten or more nucleotides, the majority of which are pyrimidines (uracil and cytosine nucleotides) (Figure 2-1)[30].

Splicing begins when the U1 snRNP binds to the 5' splice site of a pre-mRNA through complementary base pairing. This step is followed by the binding of the U2 snRNP auxiliary factor (U2AF) to the 3' splice site. U2AF is a heterodimer composed of a 65 and a 35 kDa subunit (U2AF65 and U2AF35 respectively). The U2AF35 heterodimer recognizes and binds to the 3' AG motif and U2AF65 binds the PPT. Once U2AF is bound, it recruits the U2 snRNP, which binds onto the branch site, followed by the binding of U4/U6 and U5 snRNPs[31]. The branch site is another important motif required for splicing and is located 15 to 50 bp upstream of the 3' splice site. Although the motif is highly degenerate, it contains an important and conserved (> 90%) adenosine residue that donates its 2’ hydroxyl group in a nucleophilic attack on the phosphodiester backbone of the 5’ splice site that results in cleavage at that site and formation of a 2’,’5 phosphodiester bond. A lariat structure is formed as a result of the cleavage, which leaves a free 3’ hydroxyl group at the 5’ splice site that participates in a second transesterification reaction with the guanosine residue at the 3’ splice. This two-step reaction ultimately joins the two exons with one another, resulting in the removal of the intron[30].

2.2.2 Spliceosome Interactions with pre-mRNA

Many of the spliceosomal contacts with pre-mRNA consist of RNA:RNA interactions that have been determined through the use of several techniques, such as mutagenesis of snRNAs and photo cross-linking[32,33]. For instance, the 5’ splice site is known to base pair with two snRNPs – U1 and U6. Initially, the U1 snRNP binds to the intron at positions 1, 2, 3, 5, and 6 of the splice site as well as positions -1 and -2 of the exon relative to the 5’ splice site[34]. During later stages of splicing, U1 dissociates and U6 comes in to create new base pair interactions with the 5’ splice site at positions 4, 5, and 6[34,35]. U5 joins the 5’ and 3’ splice sites through simultaneous contacts at positions -1, -2, and -3 of the exon relative to the 5’ splice site and positions 1 and 2 of the exon relative to the 3’ splice site[34,36].

While many of the previously described base pair contacts are precise in terms of sequence and sequence length, the binding affinity of U2AF to the polypyrimidine tract depends on not only the pyrimidine content but also on the length of the polypyrimidine tract[34]. For this reason, there is no exact sequence or sequence length within pre-mRNAs that bind to the U2AF, although the U2AF35 subunit does recognize and bind to the AG motif in the 3' splice site at positions -1 and -2 as well as position +1 in the exon[37].

2.3 Results

2.3.1 Genetic Variation by Gene Region

The measured values of genetic variation for various genomic sections are similar to those from other studies[12,17,18,38], particularly those from *Halushka et al* (Table 2-1). As expected, synonymous coding sites show the highest θ and π values, supporting our assumption of functional neutrality for these sites; by comparison, the nonsynonymous coding sites show greatly reduced diversity ($t=18.91$, $P<.001$ for θ; $t=13.85$, $P<.001$ for π), consistent with their functional significance. Variation in the UTRs is also significantly reduced compared to synonymous coding variation, both for 5' UTRs ($t=2.18$, $P<.05$ for θ; $t=2.79$, $P<.01$ for π) and 3' UTRs ($t=3.93$, $P<.001$ for θ; $t=3.48$, $P<.001$ for π).

Table 2-1 Population genetic parameters θ and π (± s.e.) derived from the combined datasets.

Section	Section Details	bp	θ*	π*
Coding	All	1,336,617	7.52±0.21	4.78±0.23
	Nonsynonymous	1,029,502	5.41±.0.23	3.04±0.24
	Synonymous	307,115	14.49±0.44	10.55±0.51
Intron	5', positions +1 to +50	431,348	11.25±0.34	7.89±0.37
	5', positions +7 to +50	378,756	12.29±0.37	8.60±0.42
	5', positions +1 to +6	52,592	3.75±0.46	2.73±0.59
	5', positions +1 to +6 (3 most proximal)	11,106	4.27±0.87	2.93±1.12
	5', positions +1 to +6 (3 most distal)	11,156	2.67±0.64	1.46±0.70
	3', positions -1 to -50	431,049	10.53±0.33	7.53±0.35
	3', positions -7 to -50	378,403	11.10±0.35	8.14±0.39
	3', positions -1 to -6	52,646	6.46±0.70	3.14±0.52
	3', positions -1 to -6	11,148	8.72±1.22	4.78±1.09

	(3 most proximal)			
	3’, positions -1 to -6 (3 most distal)	11,172	6.17±0.99	3.28±0.91
UTR	5' (positions +1 to +50)	42,981	12.34±0.89	7.98±0.83
	3' (positions -1 to -50)	35,953	10.85±0.72	7.13±0.79

*** Values of θ and π are x10^4. Total of 941 genes; however, number of genes varies within each section.**

Of particular interest, intronic variation for the first 50 nucleotides is significantly reduced compared to synonymous coding sites, both on the 5’ side (t=5.86, P<.001 for θ; t=4.03, P<.001 for π) and the 3’ side (t=7.16, P<.001 for θ; t=4.76, P<.001 for π). It is clear that much of this reduction is due to positions +1 to +6 on the 5’ side (t=16.99 for θ, P<.001; t=9.92, P<.001 for π) and positions -1 to -6 on the 3’ side (t=9.54, P<.001 for θ; t=9.86, P<.001 for π). However, the remaining positions also show significant reduction in variation compared to synonymous sites, both for positions +7 to +50 on the 5’ side (t=3.65, P<.001 for θ; t=2.68, P<.01 for π) and positions -7 to -50 on the 3’ side (t=5.91, P<.001 for θ; t=4.76, P<.001 for π). We also note that while there is slightly reduced (non-significant) variation on the 3’ side of introns compared to the 5’ side overall, the pattern is quite different comparing the first 6 nucleotides versus the remaining 44. For the first 6 nucleotides (positions +1 to +6 on the 5’ side, -1 to -6 on the 3’ side), there is less variation on the 5’ side than the 3’ side (t=3.24, P<.001 for θ; t=0.53, P=NS for π). By contrast, for the remaining nucleotides, there is less variation on the 3’ side than the 5’ side (t=2.29, P=.011 for θ; t=0.80, P=NS for π). This likely reflects greater functional constraints on the 3’ side than the 5’ side, suggesting the reduced variation of the first 6 nucleotides of the 5' splice site is due to that region harboring one of the very few mechanisms of 5' splice site recognition (elaborated in next sections below). We were also interested to establish whether the reduced variation in the first six positions on either the 5’ or 3’ side of introns varied based on the position of the intron in the gene. One might expect greater functional consequences for the transcribed gene due to mutations in the first six nucleotides of the most proximal introns compared to the most distal introns. Therefore, we compared variation at these positions for the 3 most proximal introns in genes versus the 3 most distal introns. Surprisingly, we actually observed greater genetic variation in the proximal introns compared to the distal introns, although these differences were not statistically significant. The patterns of reduced variation in introns observed here lead to the finer analysis of specific nucleotide sites, as described below.

2.3.2 5' Splice Site

The measured genetic variation observed within each individual position in the intronic 5’ splice site highlights important functional information that agrees with some of the known interactions that take place in this region (Figure 2-2 and Figure 2-3). Positions +1 through +6 are strongly reduced for both θ and π values (Figure 2-2). These reductions are highly statistically significant (Figure 2-3) reinforcing the known interactions with the U1 and U6 snRNPs described previously (2.2.2). A second region that appears to have reduced genetic variation is located at positions +24 through +30. This result suggests the presence of a distance-dependent functional element that is consistent across introns and genes. Due to the way positions were chosen for analysis using the distance from the splice site, this general region may represent the preferred location of a functional motif consistent across all introns and genes. To further characterize the properties of this cluster, a sliding window of length 6 basepairs was used to measure the genetic variation across every hexamer within this region (2.5). This allowed the measurement of the joint statistical significance for the individual positions within the predicted functional cluster in addition to all neighboring areas and effectively "smoothes" the observed distribution. Figure 2-4 and Figure 2-5 confirms that both the 5' splice site and the more distal predicted functional region show statistical significance as in the original analysis (hexamers starting at positions +21 through +25 correspond to positions +21 through +30), while the remainder of the sequence shows polymorphism levels associated with nonfunctional regions. The sequence range from position +36 through +45 also shows a modest tendency toward reduced levels of variation, but these locations were not statistically significant.

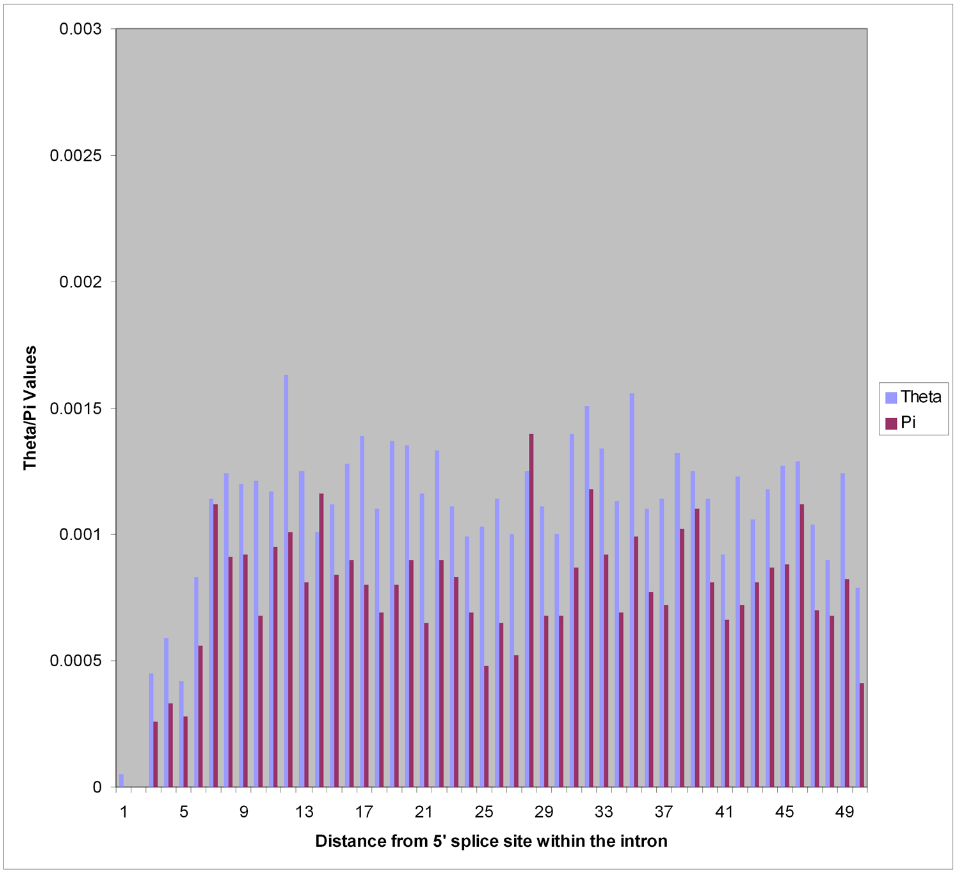

Figure 2-2 Distribution of human polymorphism for the 5' splice site using the combined datasets. θ and π values are shown for each nucleotide position.

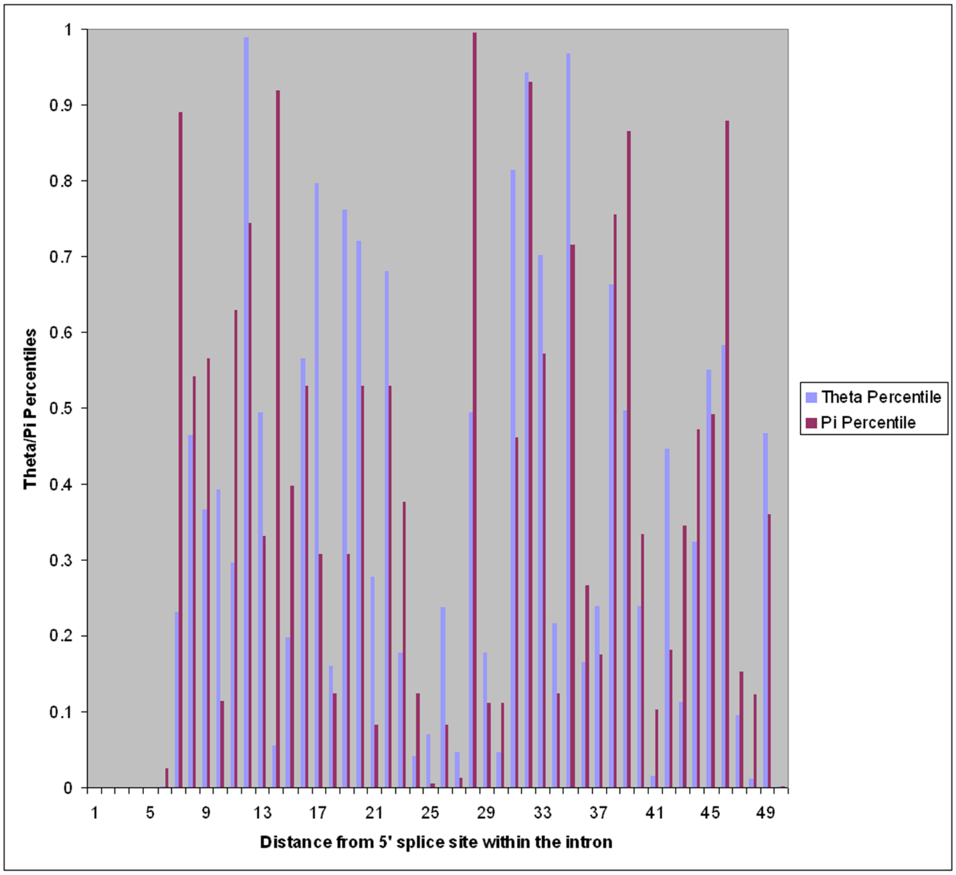

Figure 2-3 Distribution of human polymorphism for the 5' splice site using the combined datasets. Percentiles are shown for each nucleotide position.

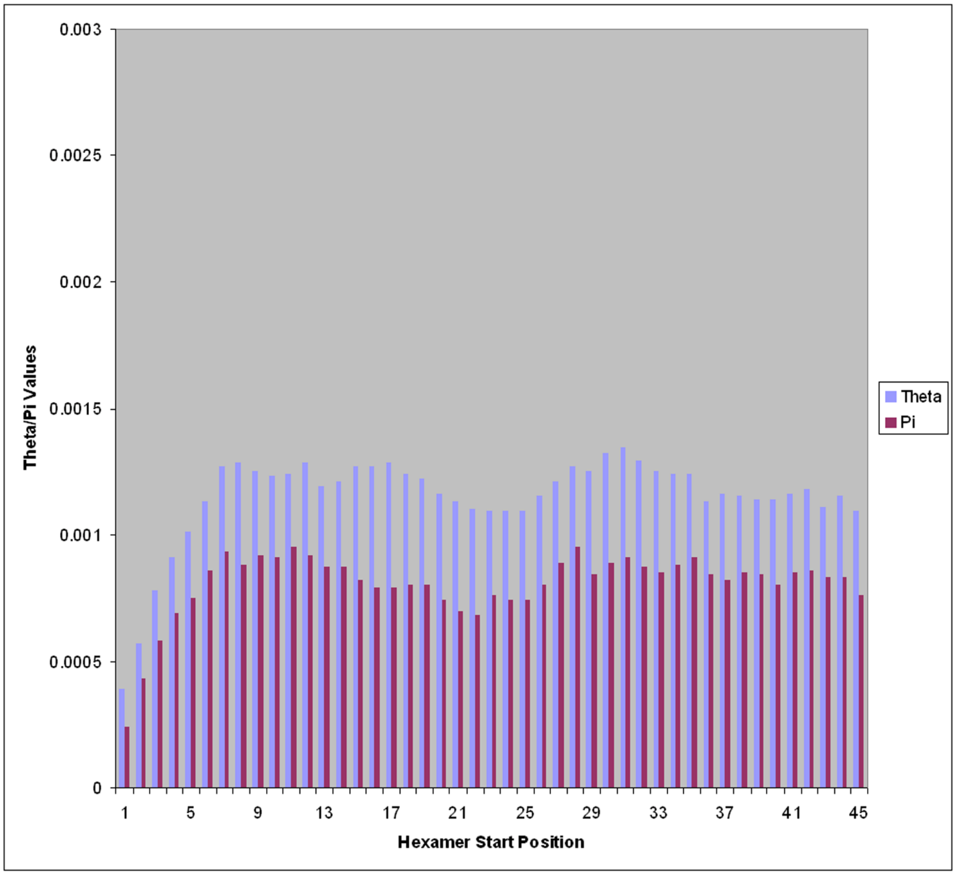

Figure 2-4 Distribution of human polymorphism for hexamers in the 5’ splice site using the combined datasets. θ and π values are shown for each hexamer starting at the shown nucleotide position.

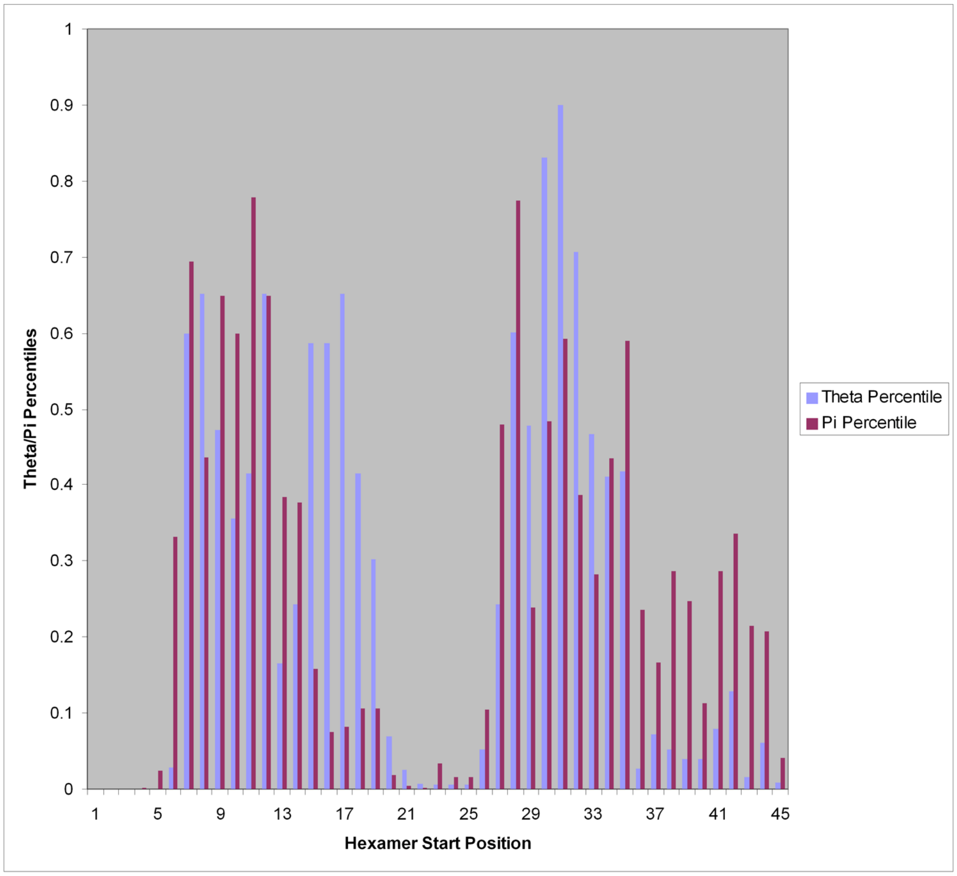

Figure 2-5 Distribution of human polymorphism for hexamers in the 5' splice site using the combined datasets. Percentiles are shown for each hexamer starting at the shown nucleotide position.

The same analyses described above were performed for the individual datasets to look for confirmation that the signals observed in Figure 2-2 and Figure 2-4 are generalized properties of all introns and genes. The results using only the NIEHS phase 1 data, the NIEHS phase 2 data, the Seattle SNP data, and the PMT data (2.5) are shown in Figure 2-6 through Figure 2-13. NIEHS phase 1 and 2, Seattle SNP, and PMT data all show similar results for nucleotides +1 to +6. Of notable importance, positions +1 and +2, the two most conserved positions of the 5' splice site, show statistically reduced θ and π values in every dataset. Positions +3 through +6 are also statistically significant across most datasets, again agreeing with the known binding positions with the U1 and U6 snRNPs. The only other clusters that appear reduced across the individual datasets are from positions +24 through +27, as found in the combined data, although the overall significance of the results in each of the individual datasets is not as prominent as that from the combined datasets. This is likely attributable to the decrease in the sample sizes of the individual datasets.

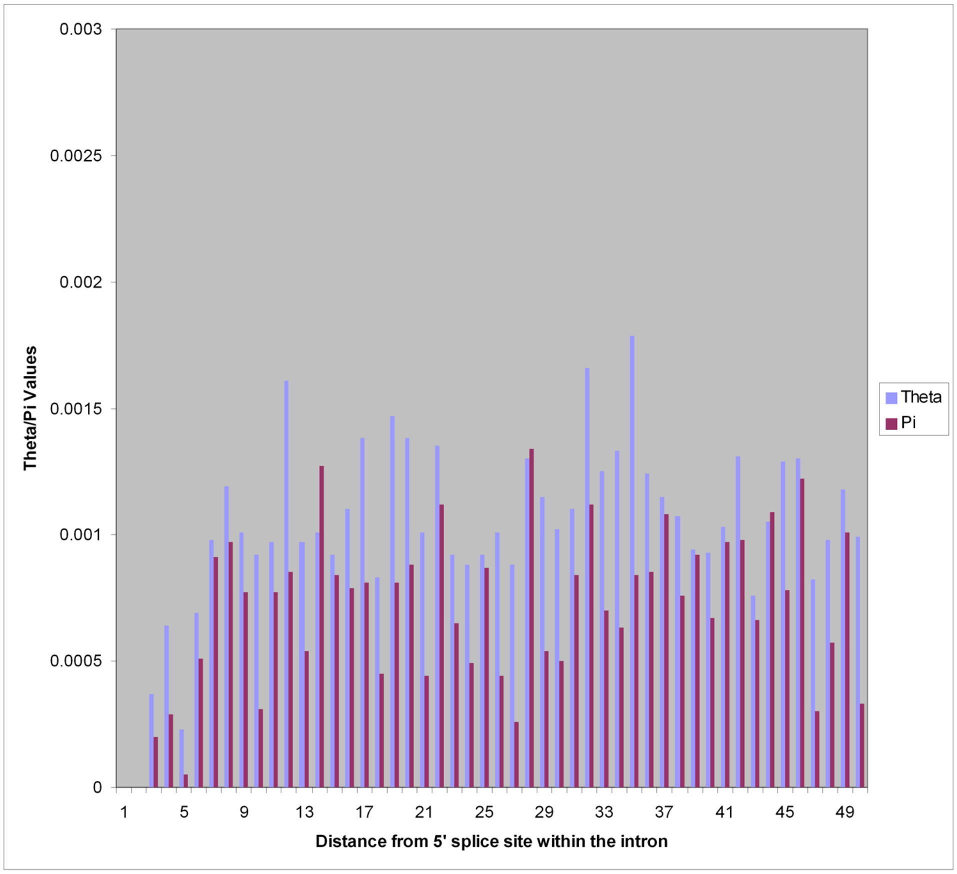

Figure 2-6 Distribution of human polymorphism for the 5' splice site using the NIEHS phase 1 dataset. θ and π values are shown for each nucleotide position.

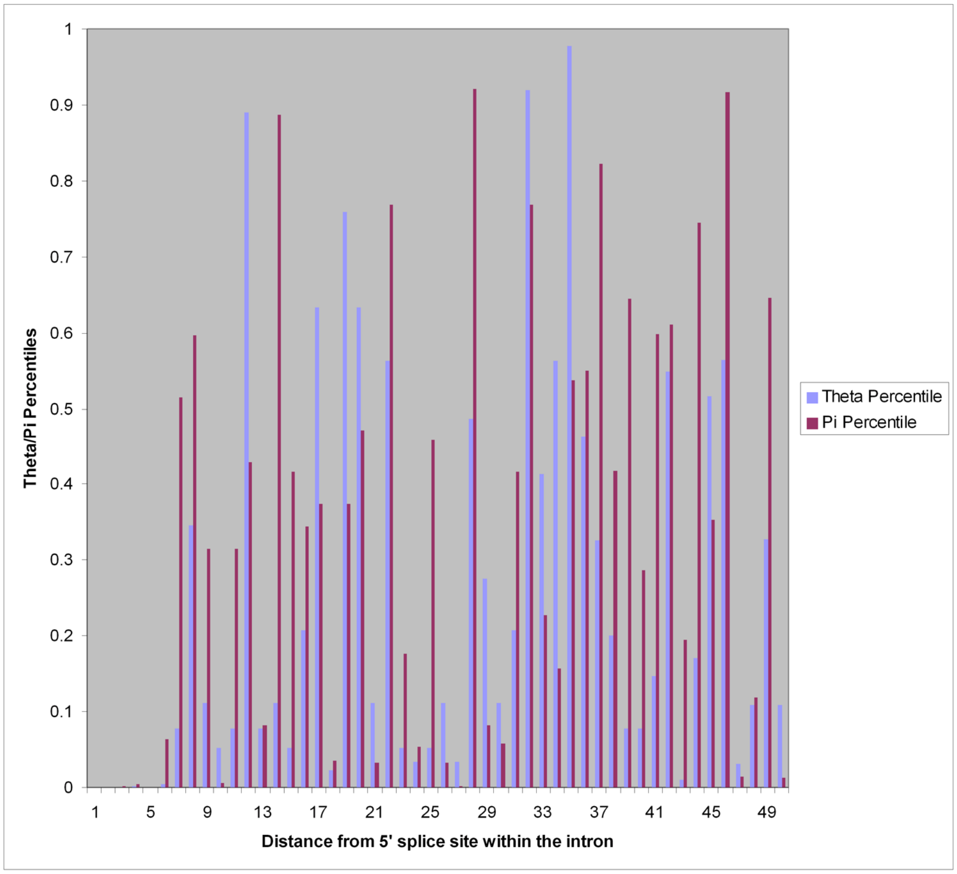

Figure 2-7 Distribution of human polymorphism for the 5' splice site using the NIEHS phase 1 dataset. Percentiles are shown for each nucleotide position.

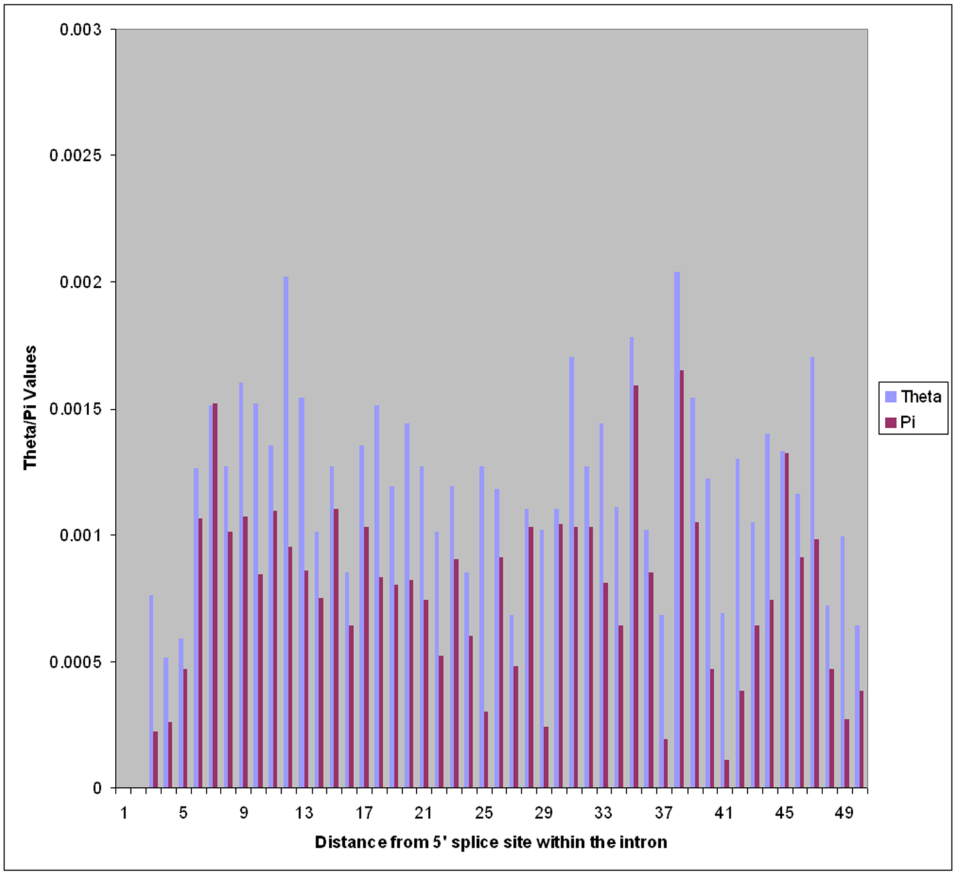

Figure 2-8 Distribution of human polymorphism for the 5' splice site using the NIEHS phase 2 dataset. θ and π values are shown for each nucleotide position.

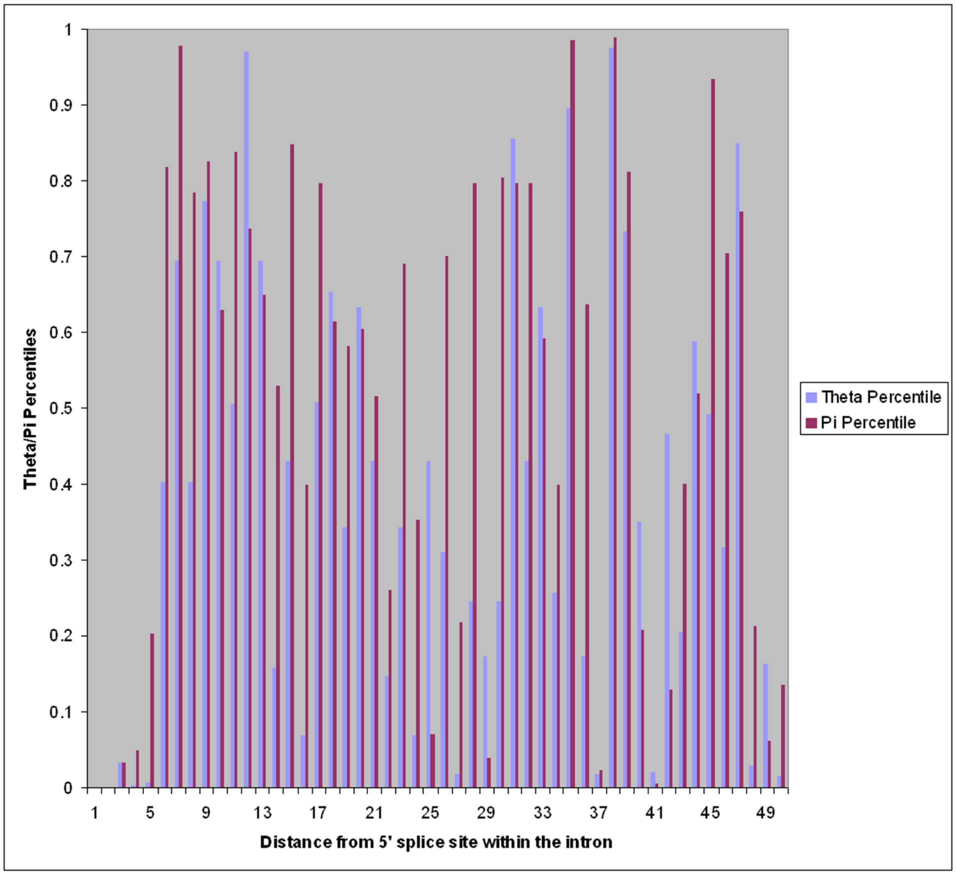

Figure 2-9 Distribution of human polymorphism for the 5' splice site using the NIEHS phase 2 dataset. Percentiles are shown for each nucleotide position.

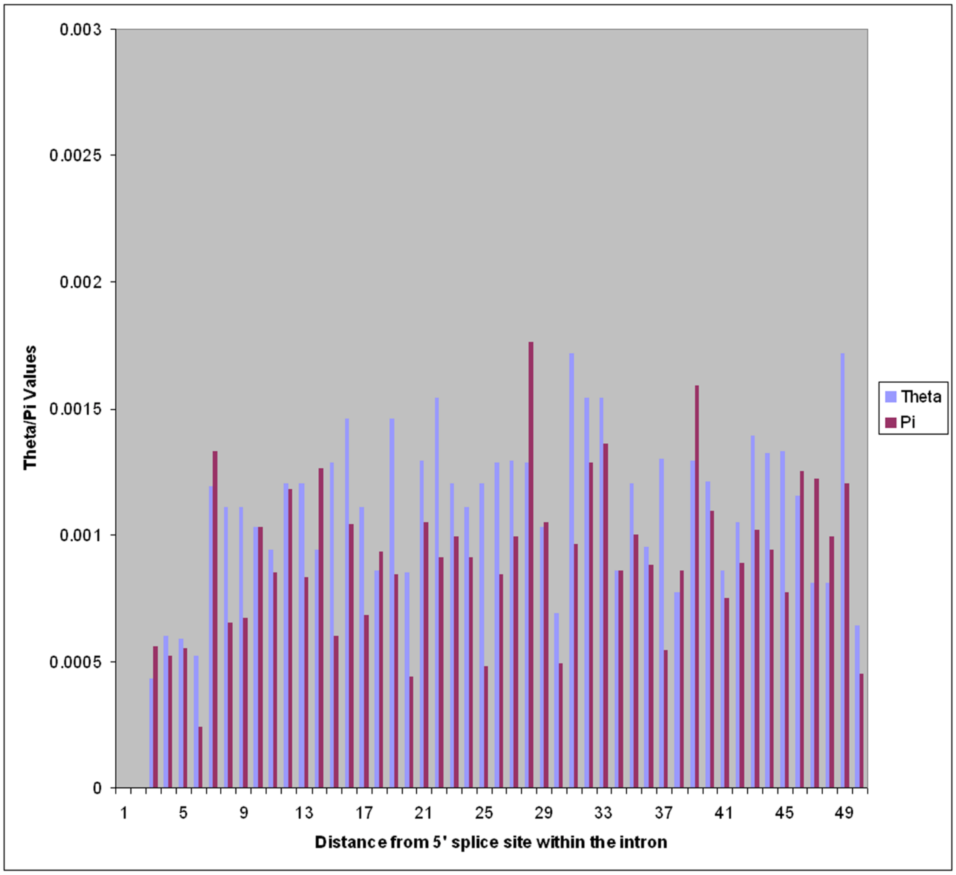

Figure 2-10 Distribution of human polymorphism for the 5' splice site using the Seattle SNP dataset. θ and π values are shown for each nucleotide position.

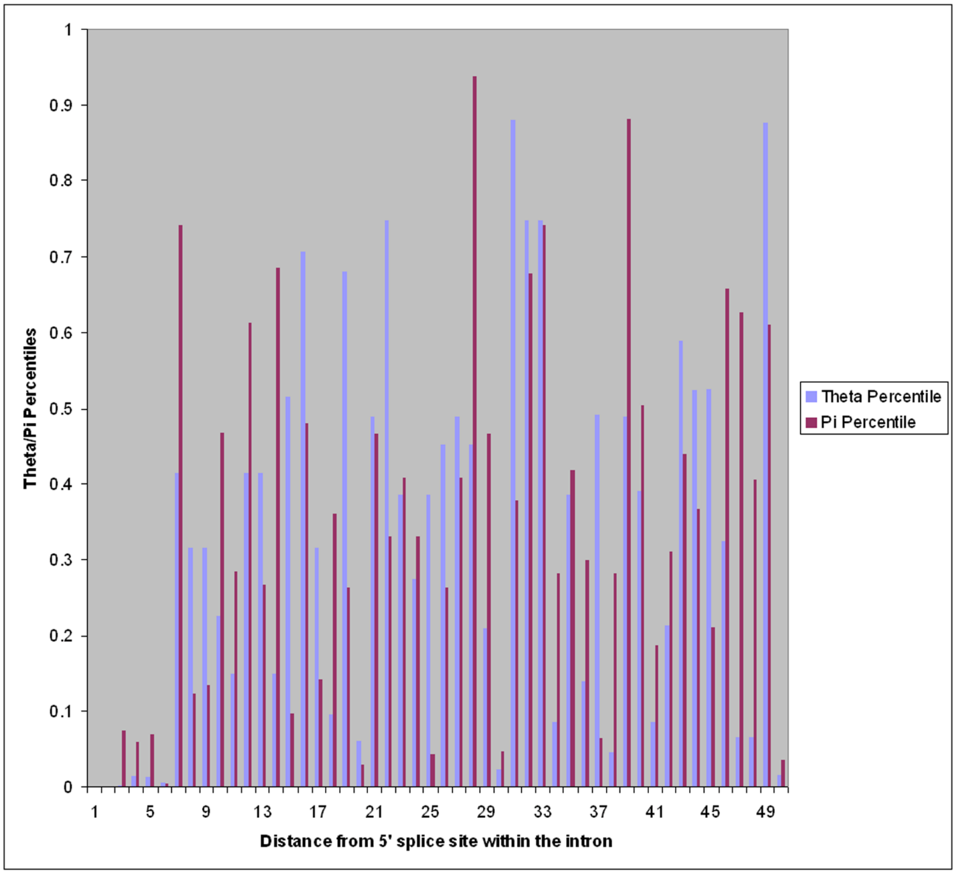

Figure 2-11 Distribution of human polymorphism for the 5' splice site using the Seattle SNP dataset. Percentiles are shown for each nucleotide position.

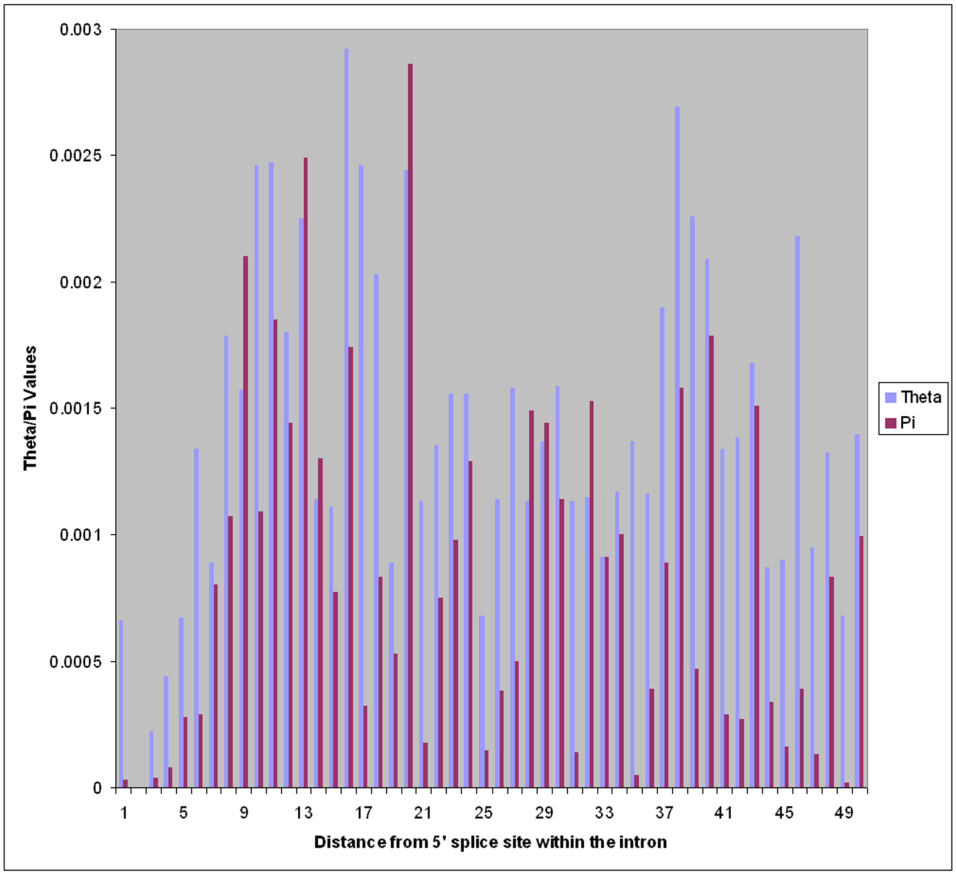

Figure 2-12 Distribution of human polymorphism for the 5' splice site using the PMT dataset. θ and π values are shown for each nucleotide position.

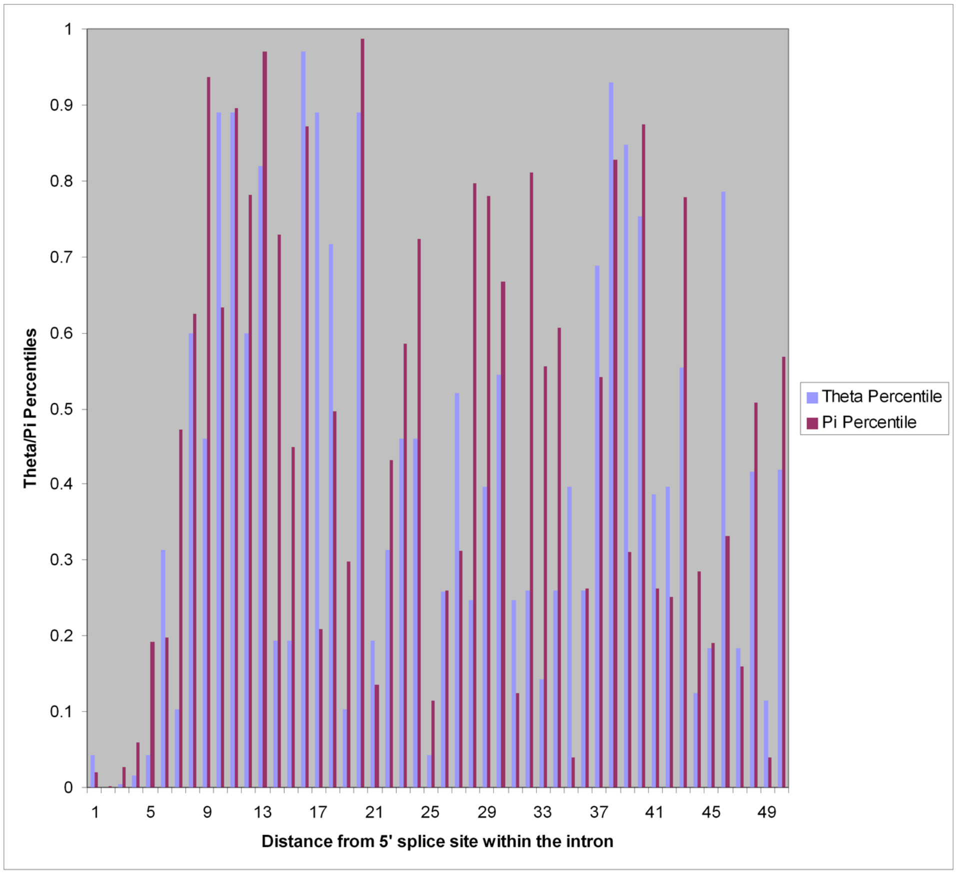

Figure 2-13 Distribution of human polymorphism for the 5' splice site using the PMT dataset. Percentiles are shown for each nucleotide position.

Because of known differences in genetic variation between ethnic groups, African Americans and Caucasians, the two largest groups, were also analyzed separately. Although several datasets have multiple ethnic groups, African Americans and Caucasians are the only groups across all datasets that had sufficient numbers of individuals for meaningful subset analyses. Only NIEHS phase 2, Seattle SNP, and PMT data were used in these analyses since these were the only sets that contained identifiable African American and Caucasian data. Smoothed graphs for the African Americans and Caucasians are given in Figure 2-14 through Figure 2-17, respectively. Both ethnicities show significantly reduced genetic variation in the 5' splice site, as was observed in the combined dataset (Figure 2-4, Figure 2-5), although the results are more significant for the African Americans. The African American graphs show that the hexamers starting at positions +20 through +25 are significantly reduced in variation, similar to what was observed in the combined dataset. The Caucasian graphs show a similar decrease in genetic variation from positions +17 through +22 (Figure 2-16), although only the hexamer starting at position +20 is formally significant (Figure 2-17). The general trends in reduced variation are quite similar between the ethnic groups and match those from the combined datasets using all ethnic groups.

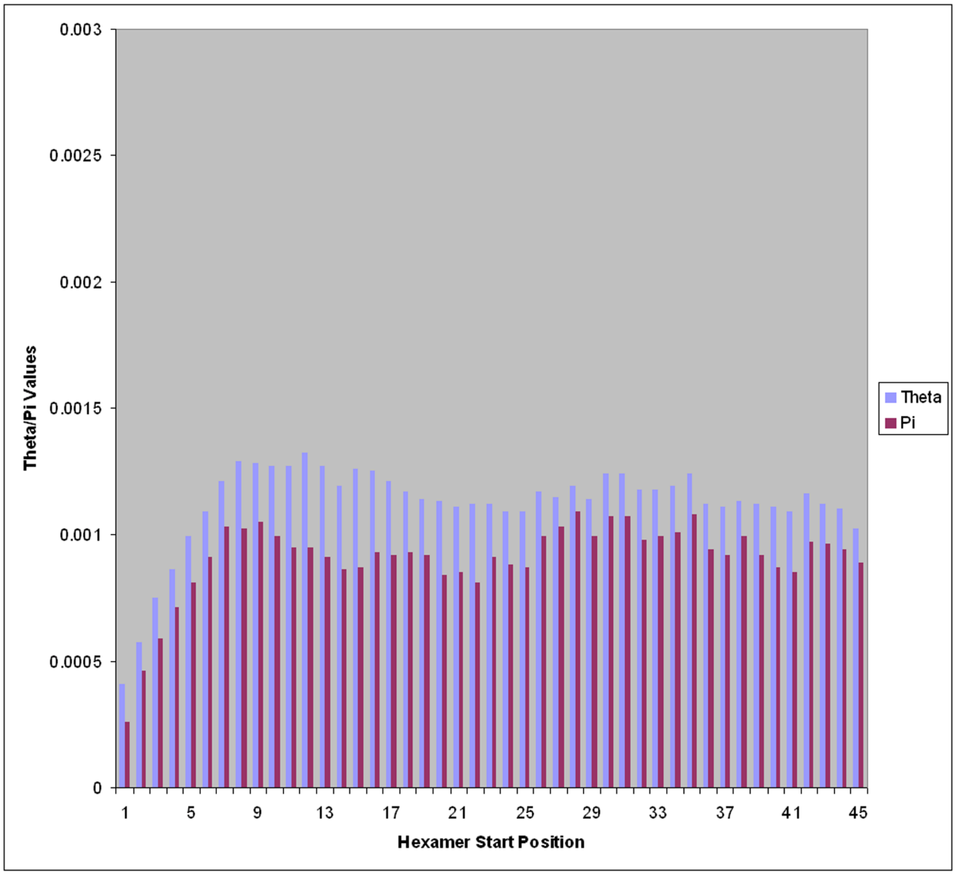

Figure 2-14 Distribution of human polymorphism for hexamers in the 5' splice site using the combined datasets and only African Americans. θ and π values are shown for each hexamer starting at the shown nucleotide position.

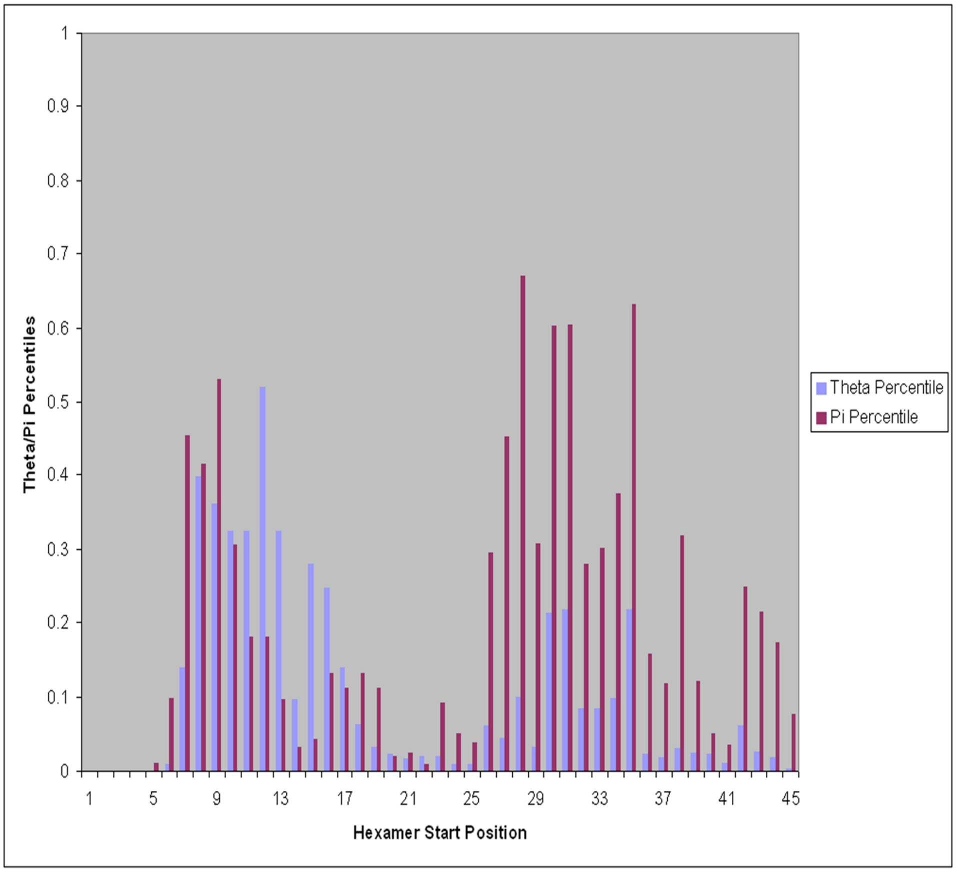

Figure 2-15 Distribution of human polymorphism for hexamers in the 5' splice site using the combined datasets and only African Americans. Percentiles are shown for each hexamer starting at the shown nucleotide position.

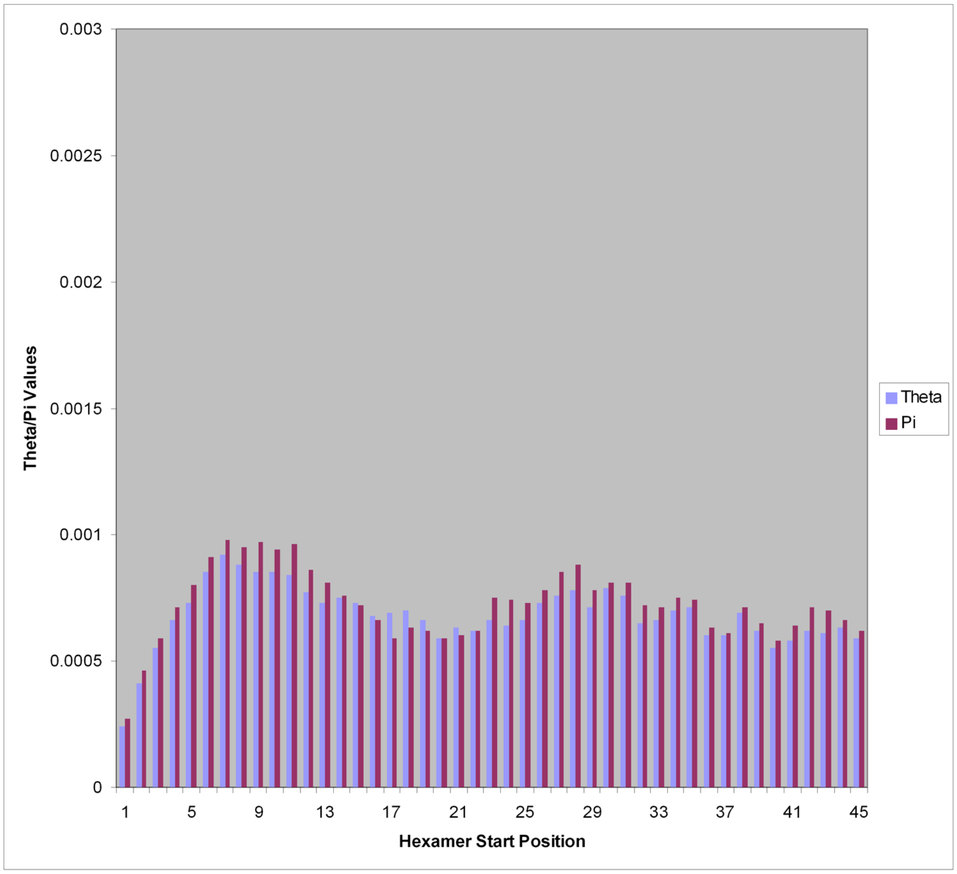

Figure 2-16 Distribution of human polymorphism for hexamers in the 5' splice site using the combined datasets and only Caucasians. θ and π values are shown for each hexamer starting at the shown nucleotide position.

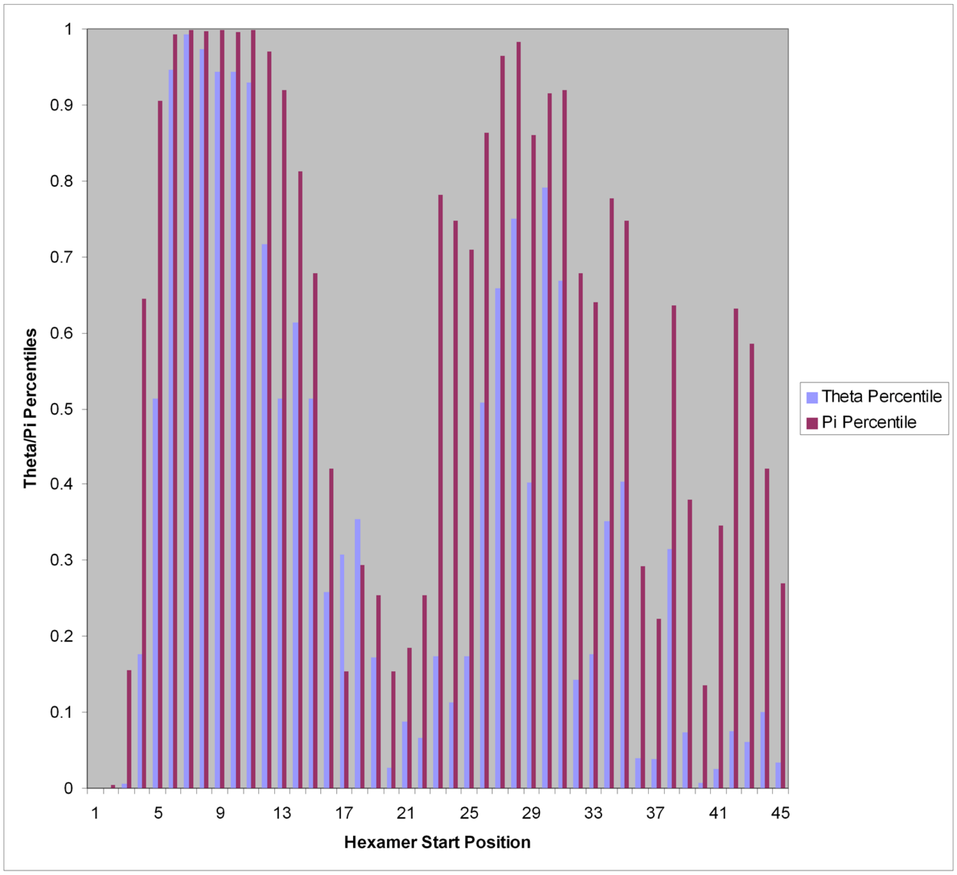

Figure 2-17 Distribution of human polymorphism for hexamers in the 5' splice site using the combined datasets and only Caucasians. Percentiles are shown for each hexamer starting at the shown nucleotide position.

2.3.3 3' Splice Site

Analyses similar to those described above for the 5' splice site were also performed for the 3' splice site. In this case, positions -50 through -1 were analyzed for the combined datasets and ethnicities (Figure 2-18, Figure 2-19). Of notable importance, positions -1 through -5 all showed reduced genetic variation, including position -4, which is not conserved between humans and other species[39]. The reduced variation observed at positions -1 and -2 agrees with the known binding of the U2AF35 snRNP to the AG motif (2.2.2) Due to the presence of individual positions with low genetic variation scattered throughout the region upstream of the 3' splice site, a sliding window analysis was performed to characterize the joint variation of neighboring positions for all hexamers upstream of the 3' splice site, as was done previously for the 5' splice site (Figure 2-20, Figure 2-21). Unlike the 5' splice site, this region shows an extended range of sites with reduced genetic variation, which is likely due to the increased presence of functional motifs such as the polypyrimidine tract and the branch site. The extended range of hexamers with reduced polymorphism levels from starting positions -6 through -10 reflects the functional importance of the polypyrimidine tract. Likewise, the presence of two regions with hexamers of low polymorphism levels from -23 through -27 and -35 through -39 suggest the localization of functionally important sequences from nucleotide positions -23 through -32 and -35 through -44 (e.g. the branch site).

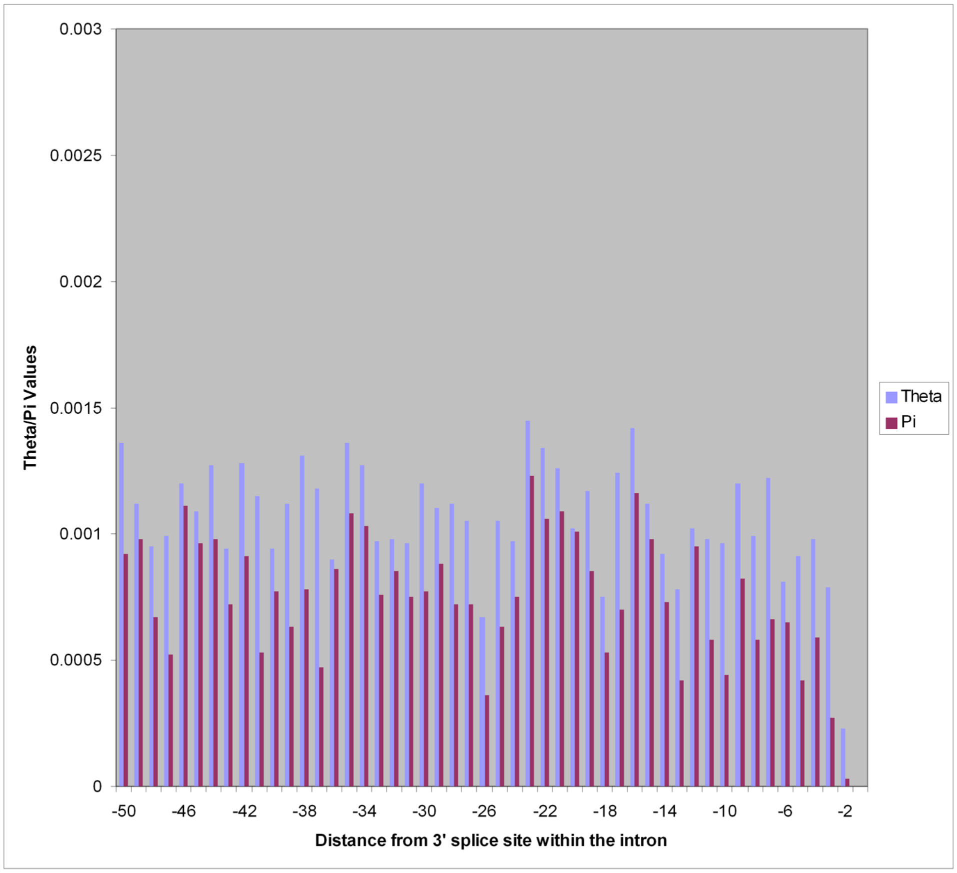

Figure 2-18 Distribution of human polymorphism for the 3' splice site using the combined datasets. θ and π values are shown for each nucleotide position.

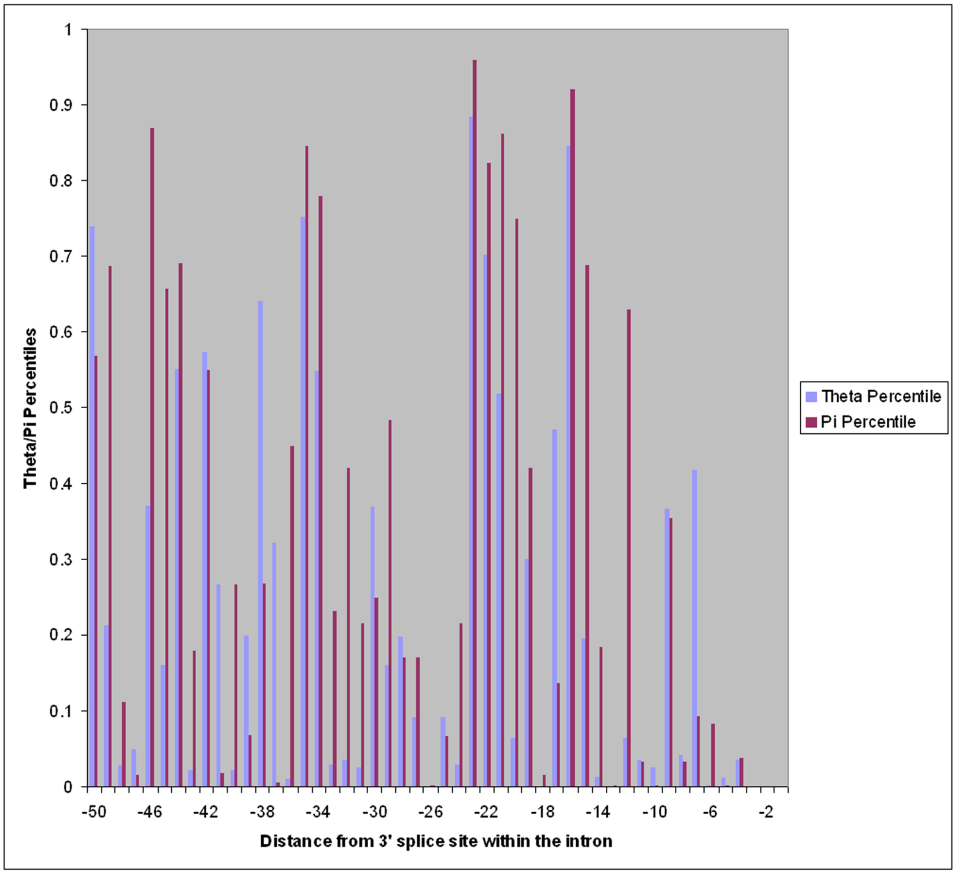

Figure 2-19 Distribution of human polymorphism for the 3' splice site using the combined datasets. Percentiles are shown for each nucleotide position.

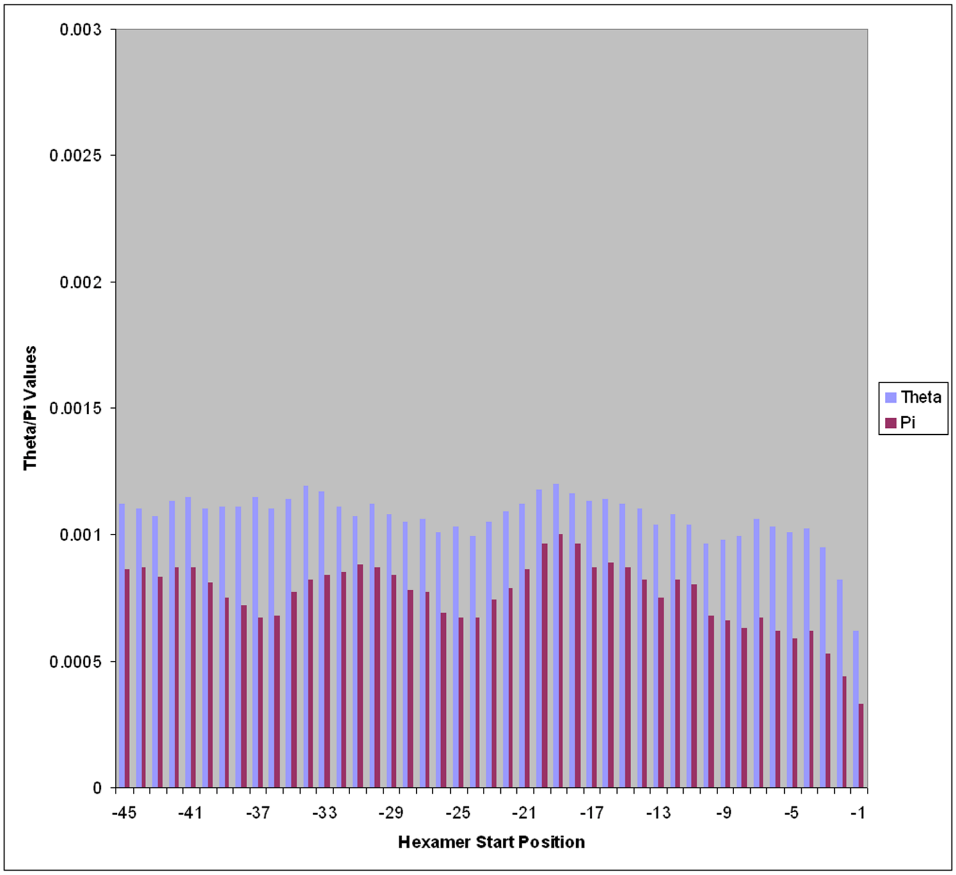

Figure 2-20 Distribution of human polymorphism for hexamers in the 3' splice site using the combined datasets. θ and π values are shown for each hexamer starting at the shown nucleotide position.

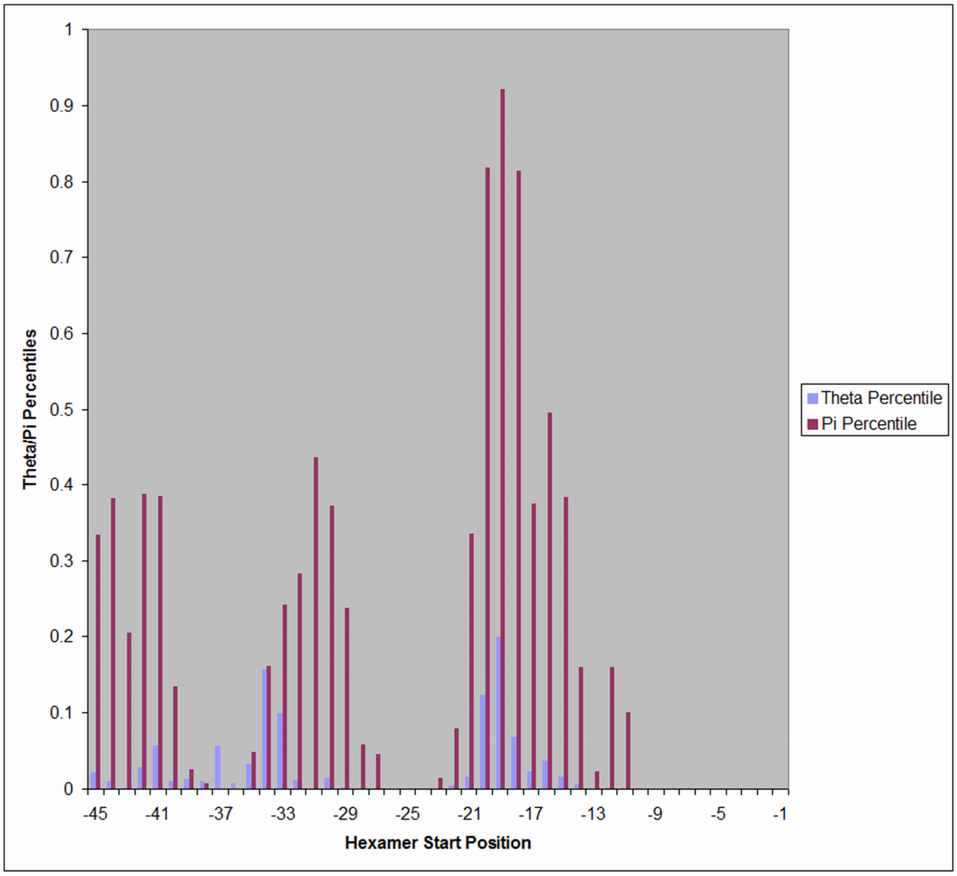

Figure 2-21 Distribution of human polymorphism for hexamers in the 3' splice site using the combined datasets. Percentiles are shown for each hexamer starting at the shown nucleotide position.

We next examined 46 intron sequences with experimentally determined branch site locations[24,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55] and found that the average position of the branch site adenosine is at position -26, which is consistent with the region from -23 through -32 harboring branch site sequences. As described earlier, branch sites can be located up to 50 nucleotides away from the 3' splice site which suggests the possibility that the region from -35 through -44 could be a second branch site localization site, although this region may also harbor a distinct, functionally important element.

We further analyzed the sequences upstream of the 3’ splice site using the NIEHS P1, P2, Seattle SNP, and PMT datasets individually (Figure 2-22, Figure 2-29). Positions -1 and -2 of the splice site consistently show statistically significant reduced genetic variation across the individual datasets, agreeing with the known binding of U2AF35 to the AG motif described previously (2.2.2). As we observed previously for the 5’ splice site, the signals are present, but less prominent for the other nucleotide positions and demonstrates the importance of a large sample size in delineating clearer patterns.

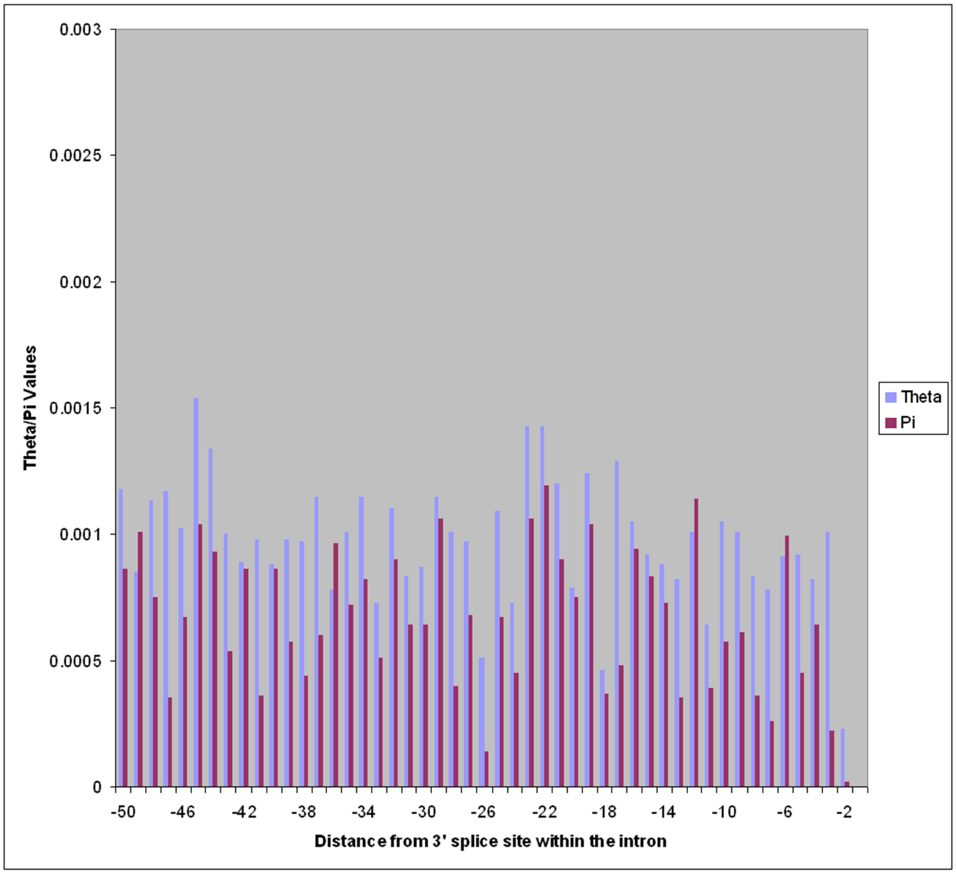

Figure 2-22 Distribution of human polymorphism for the 3’ splice site using the NIEHS phase 1 dataset. Percentiles are shown for each nucleotide position.

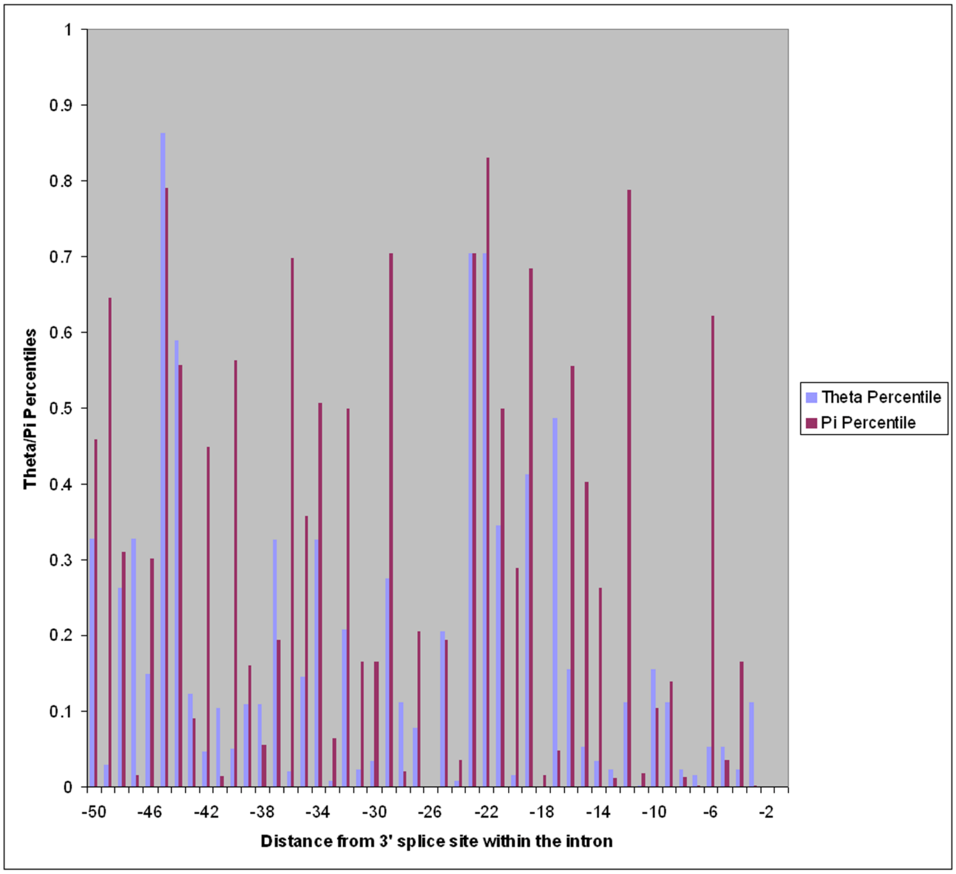

Figure 2-23 Distribution of human polymorphism for the 3' splice site using the NIEHS phase 1 dataset. Percentiles are shown for each nucleotide position.

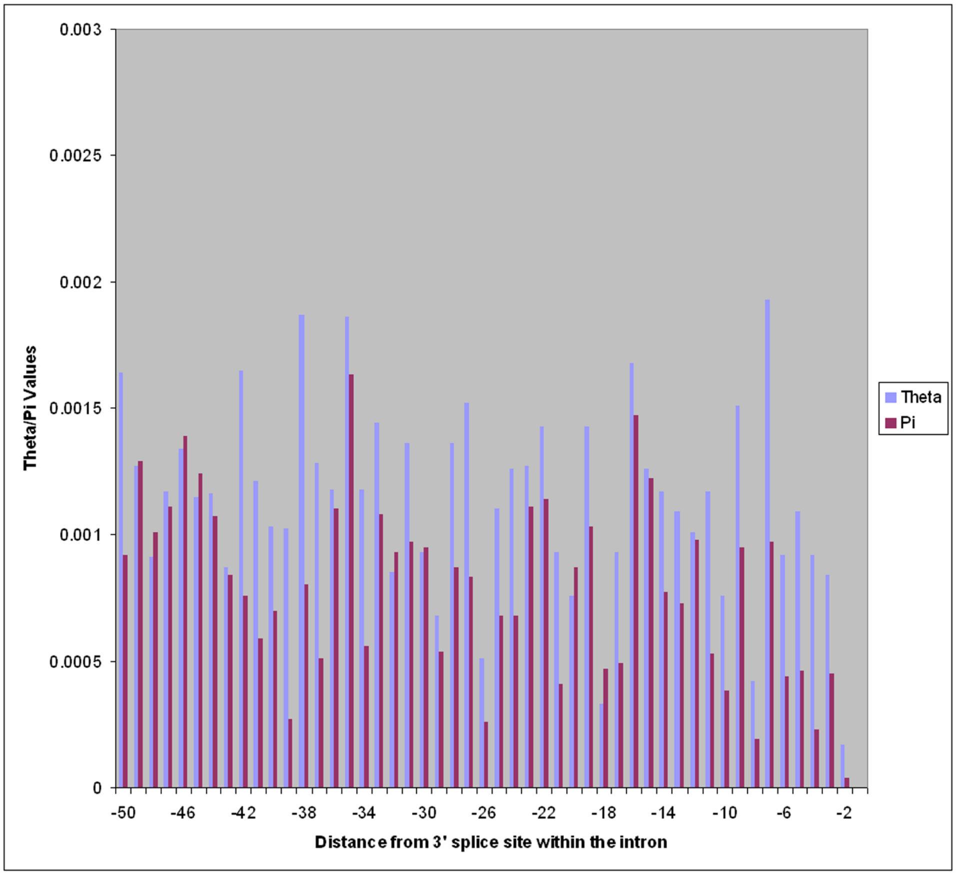

Figure 2-24 Distribution of human polymorphism for the 3' splice site using the NIEHS phase 2 dataset. θ and π values are shown for each nucleotide position.

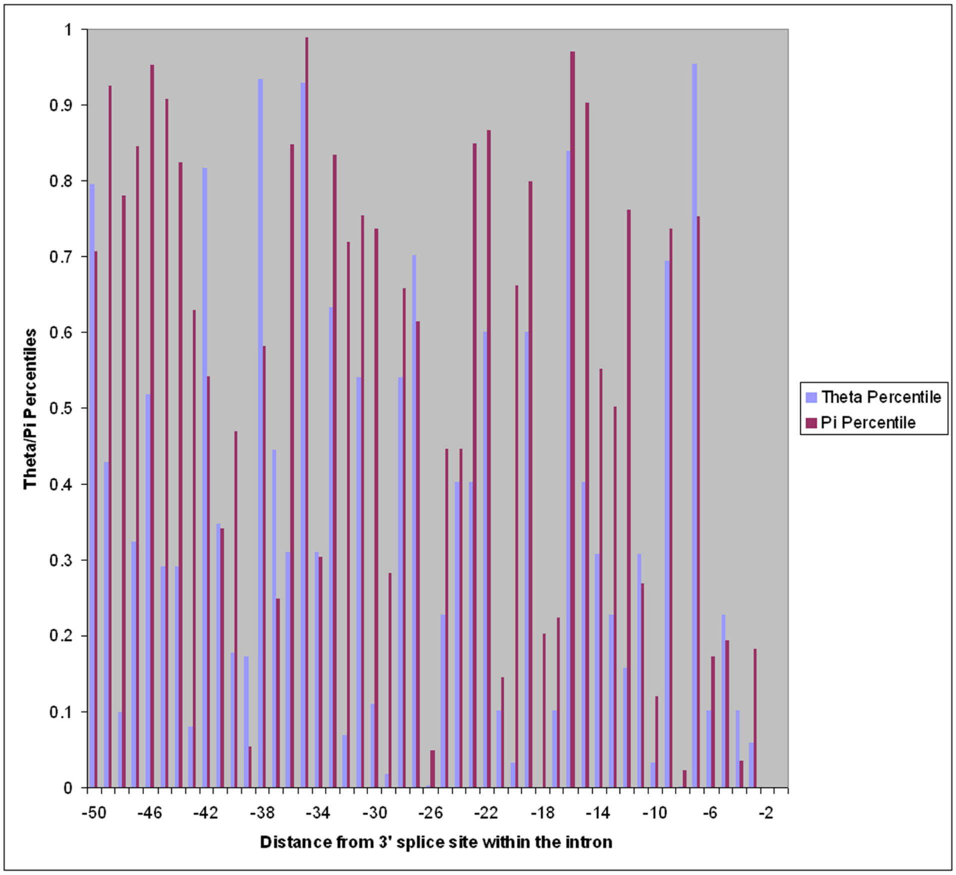

Figure 2-25 Distribution of human polymorphism for the 3’ splice site using the NIEHS phase 2 dataset. Percentiles are shown for each nucleotide position.

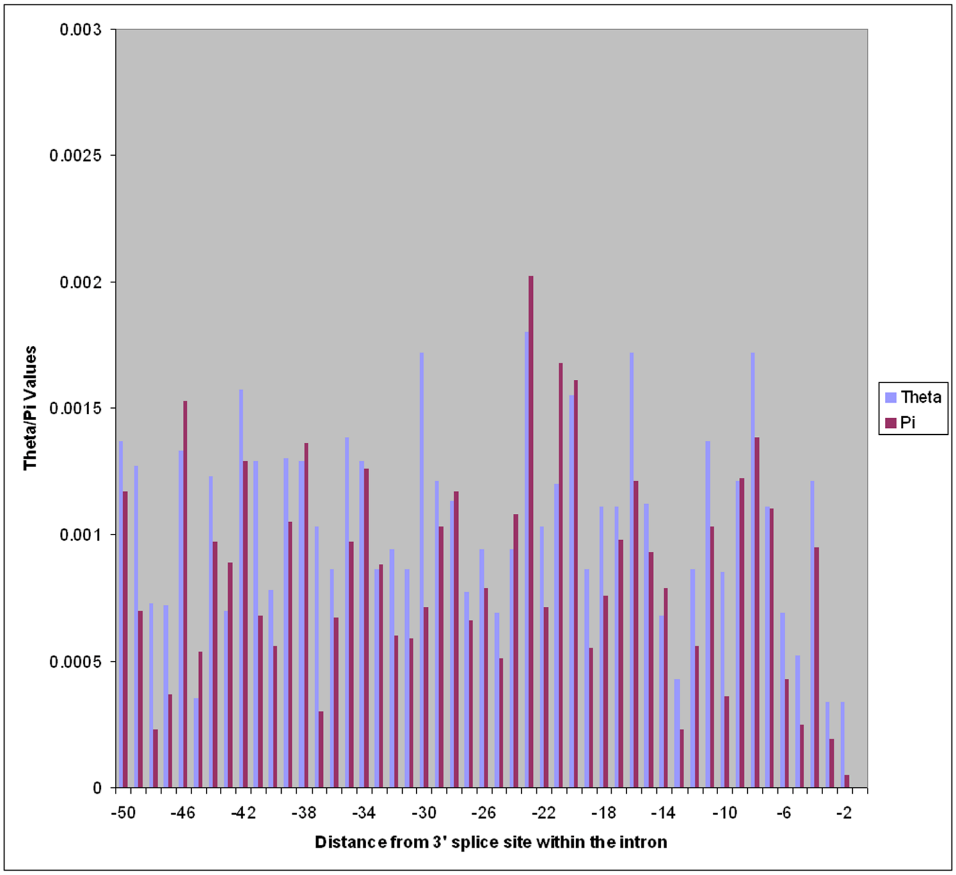

Figure 2-26 Distribution of human polymorphism for the 3’ splice site using the Seattle SNP dataset. θ and π values are shown for each nucleotide position.

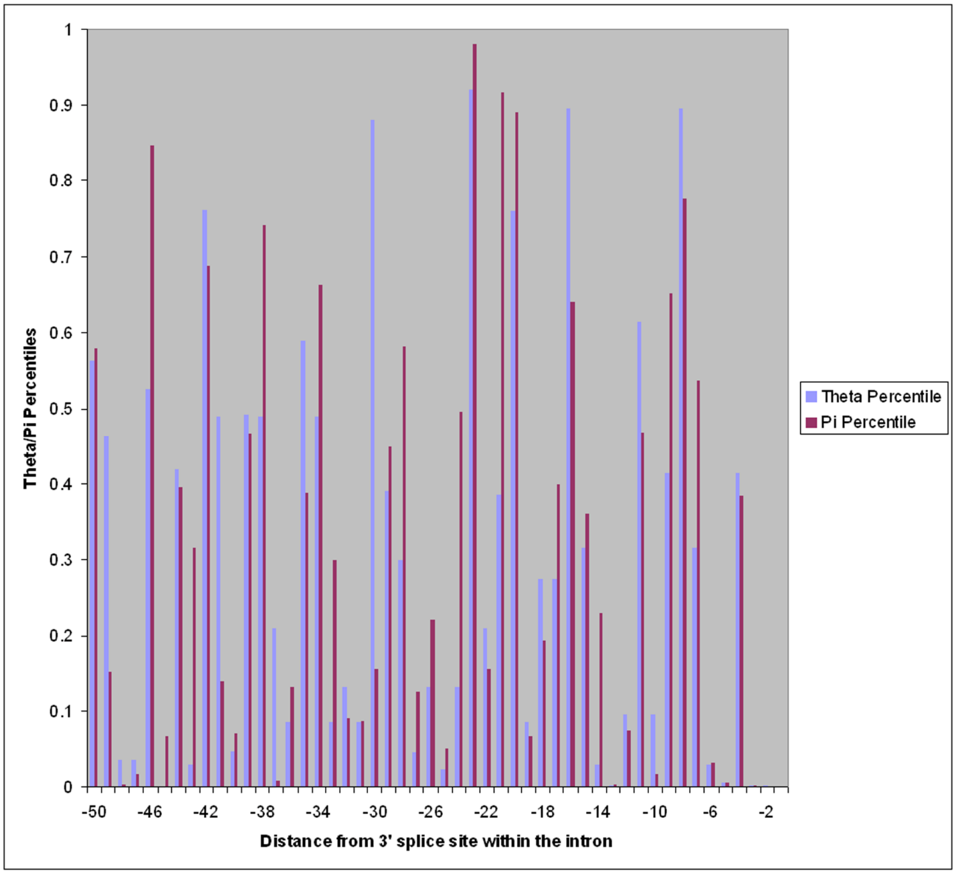

Figure 2-27 Distribution of human polymorphism for the 3’ splice site using the Seattle SNP dataset. Percentiles are shown for each nucleotide position.

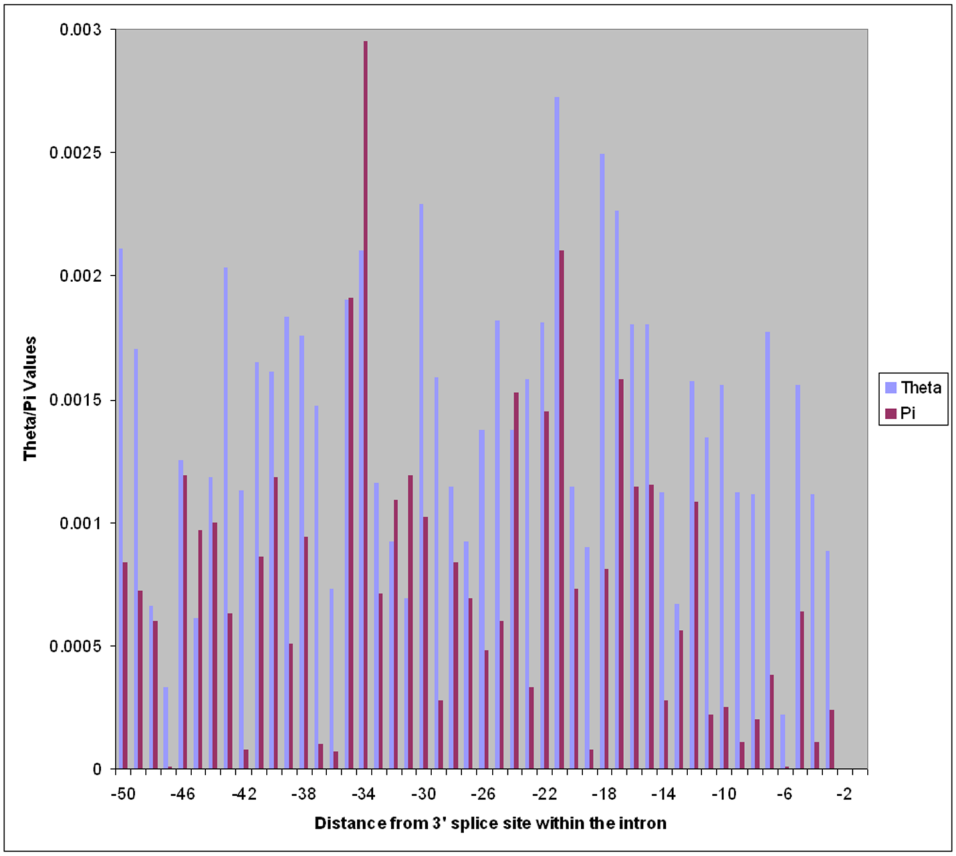

Figure 2-28 Distribution of human polymorphism for the 3' splice site using the PMT dataset. θ and π values are shown for each nucleotide position.

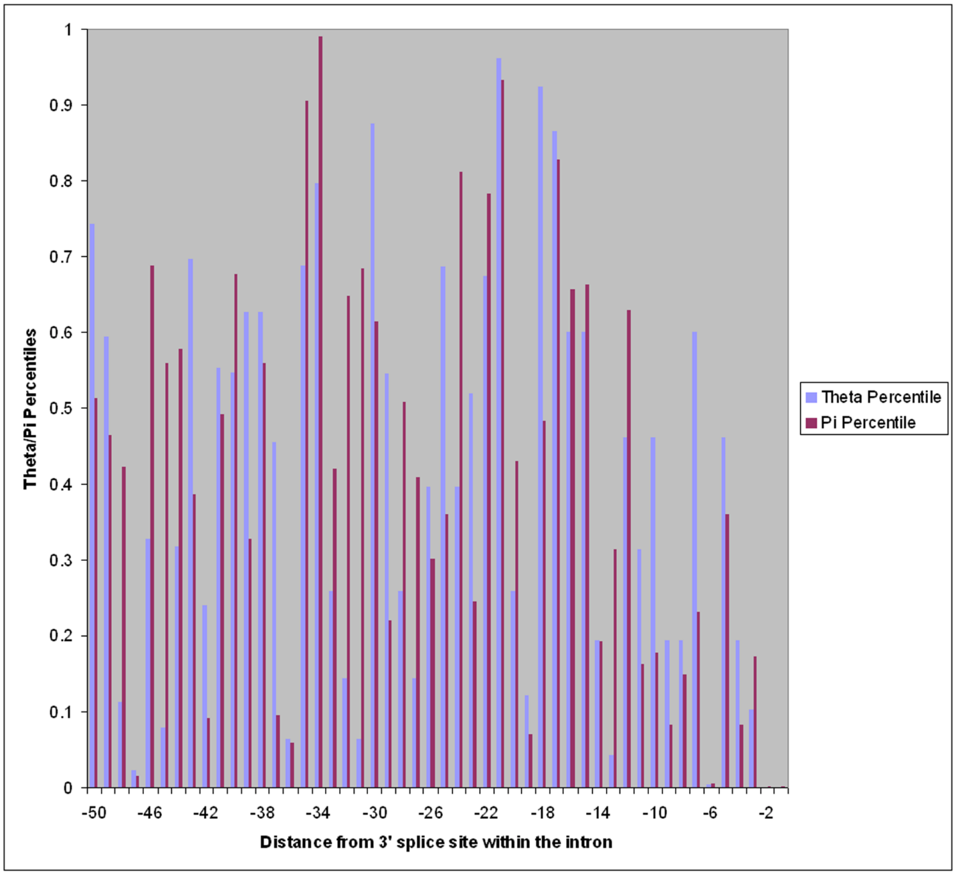

Figure 2-29 Distribution of human polymorphism for the 3' splice site using the PMT dataset. Percentiles are shown for each nucleotide position.

Smoothed graphs were generated using only African American (Figure 2-30, Figure 2-31) and Caucasian (Figure 2-32, Figure 2-33) populations from the combined datasets. The African American groups showed very similar results to the dataset generated using all ethnicities (Figure 2-20, Figure 2-21). For both regions of hexamers starting at positions -24 through -28 and -36 through -37 as well as positions -1 through -10 nearest to the 3' splice site, genetic variation was significantly reduced. The results for the Caucasian group, on the other hand, were not as prominent. While the reduced variation previously seen from hexamers starting at positions -1 to -6 and -23 to -27 are still quite evident, the region of reduced variation previously observed from positions -35 to -39 in both African Americans and the combined data is more subtle in the Caucasians alone.

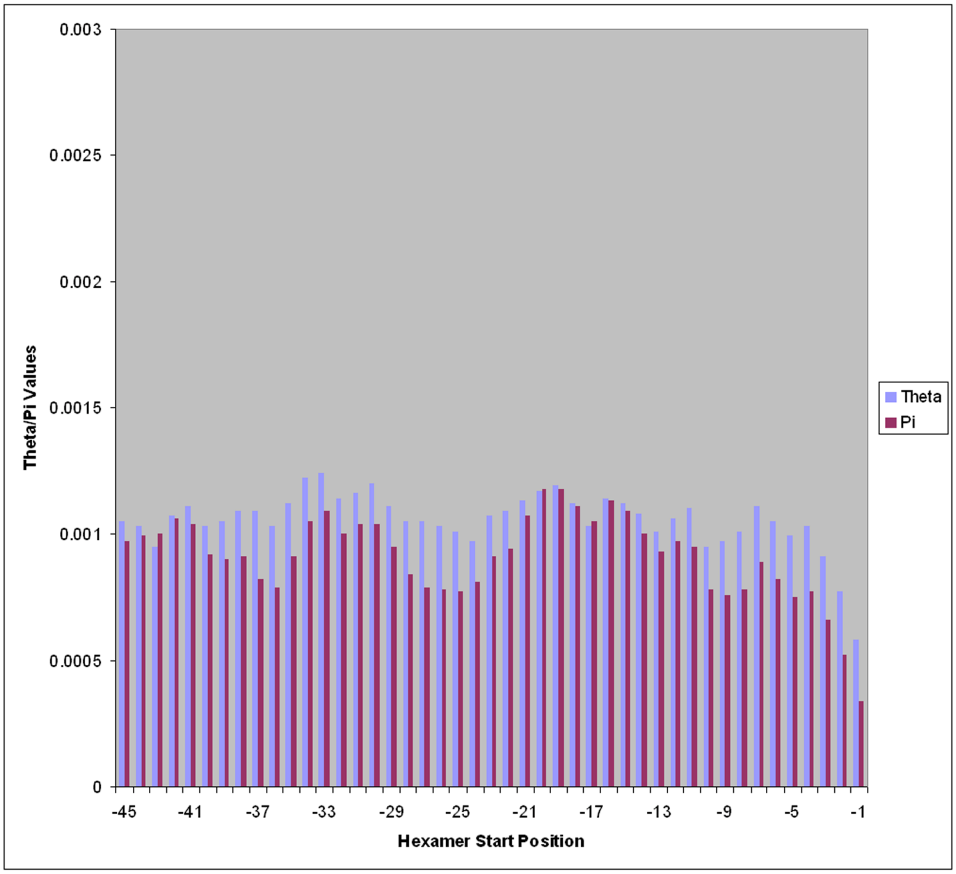

Figure 2-30 Distribution of human polymorphism for hexamers in the 3' splice site using the combined datasets and only African Americans. θ and π values are shown for each hexamer starting at the shown nucleotide position.

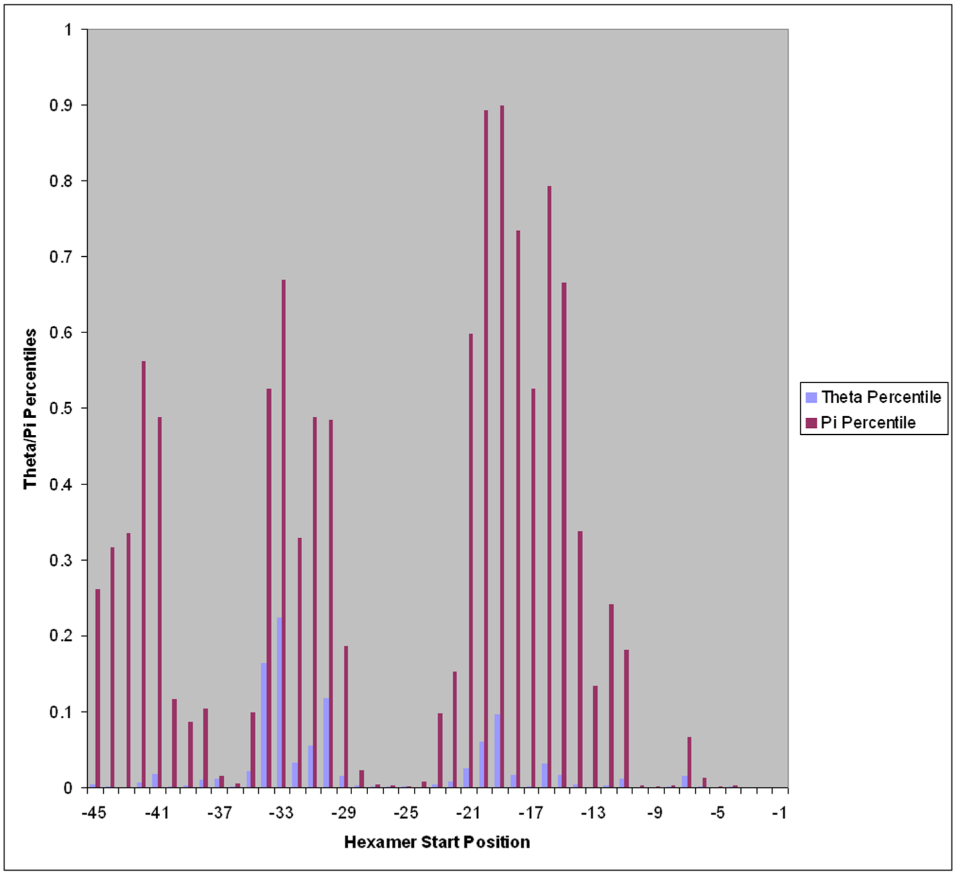

Figure 2-31 Distribution of human polymorphism for hexamers in the 3' splice site using the combined datasets and only African Americans. Percentiles are shown for each hexamer starting at the shown nucleotide position.

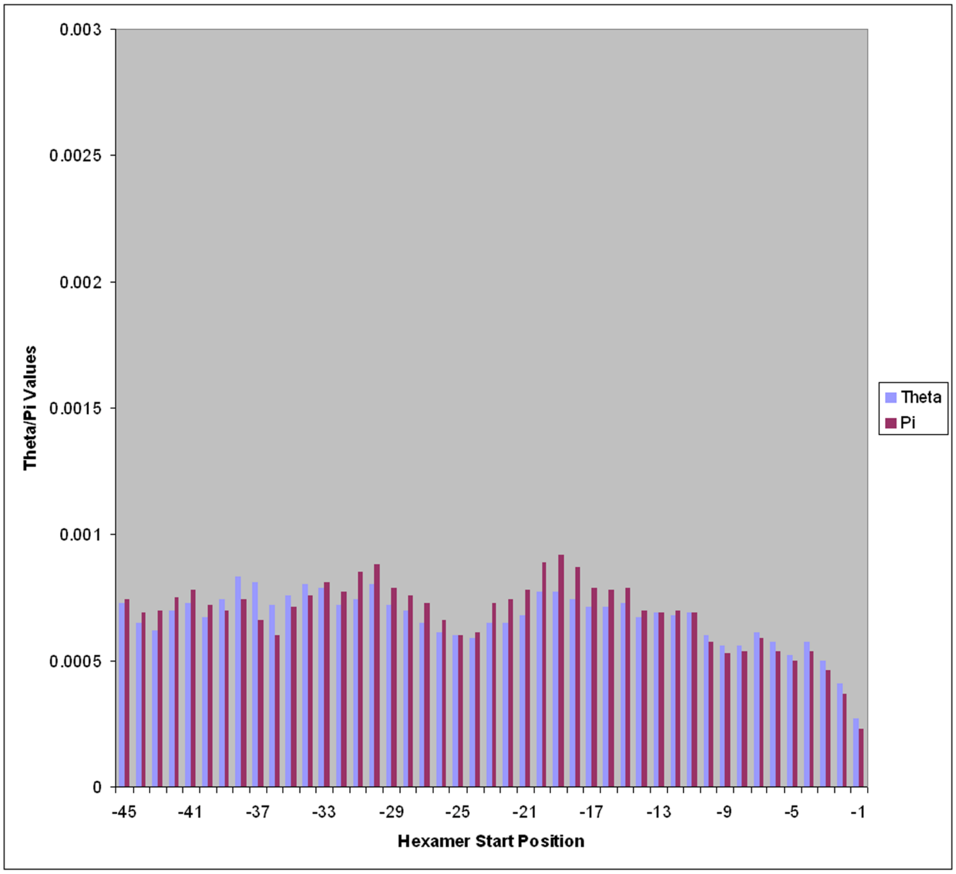

Figure 2-32 Distribution of human polymorphism for hexamers in the 3' splice site using the combined datasets and only Caucasians. θ and π values are shown for each hexamer starting at the shown nucleotide position.

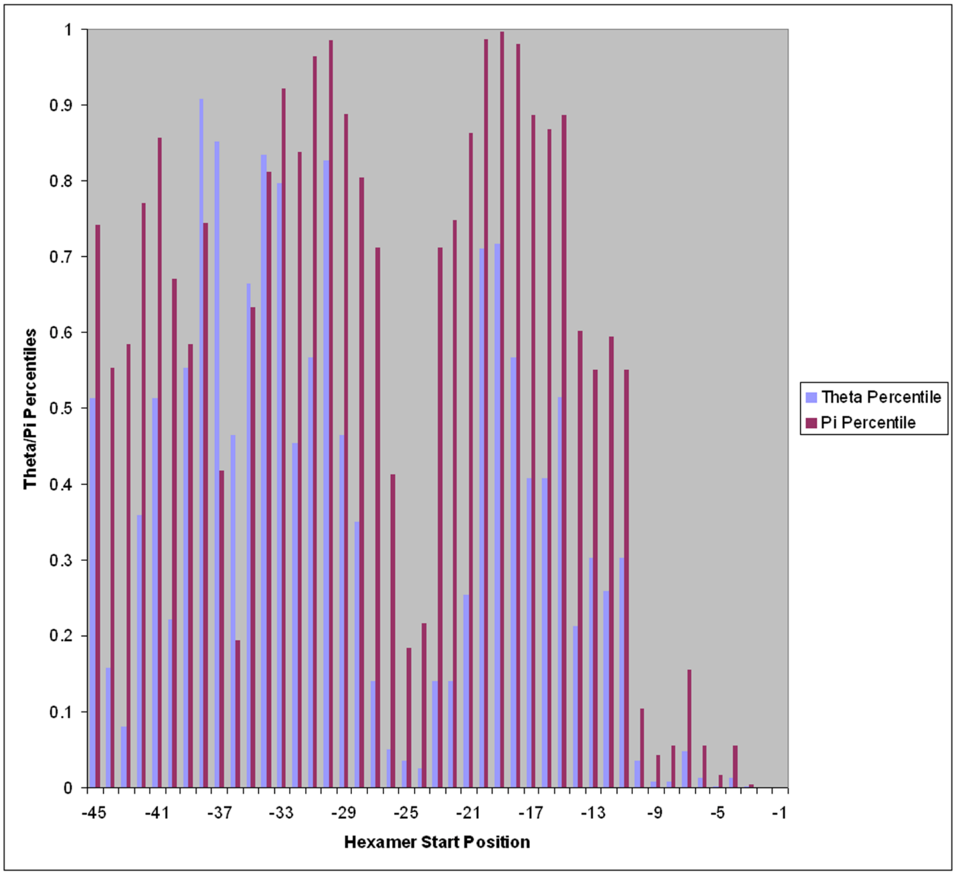

Figure 2-33 Distribution of human polymorphism for hexamers in the 3' splice site using the combined datasets and only Caucasians. Percentiles are shown for each hexamer starting at the shown nucleotide position.

2.3.4 Ethnic Variation and Tajima's D

The general observation from comparing the average θ and π values from each dataset using the 5' intronic sequence data is that African Americans (θ: 1.0×10^{-3}, 1.1×10^{-3}, 1.2×10^{-3} π:8.2×10^{-4}, 9.0×10^{-4}, 8.5×10^{-4}) have a higher polymorphism rate than Caucasians (θ: 6.2×10^{-4}, 6.4×10^{-4}, 7.5×10^{-4} π:6.4×10^{-4}, 7.0×10^{-4}, 7.7×10^{-4}) in the NIEHS phase 2, Seattle SNP, and PMT databases respectively, consistent with previous observations[17]. The same result was observed comparing the average θ and π values of the 3' intronic sequence data from the African American population (θ: 1.0×10^{-3}, 1.0×10^{-3}, 1.0×10^{-3} π:$8.8 \times 10^{-}$

^{4}, 8.9x10^{-4}, 7.9 x10^{-4}) against the Caucasian values (θ: 6.7x10^{-4}, 6.1 x10^{-4}, 7.1 x10^{-4} π:6.8x10^{-4}, 6.5x10^{-4}, 6.5 x10^{-4}) for the NIEHS phase 2, Seattle SNP, and PMT data sets respectively. The generally accepted explanation for reduced genetic variation in Caucasians is the smaller effective population size of their founders compared to those of present day Africans.

Of note, the θ and π values in Caucasians were remarkably similar, both for the 5' and 3' intronic sequence data (Figure 2-16, Figure 2-17, Figure 2-32, Figure 2-33, Table 2-2). By contrast, in African Americans, the θ values were uniformly higher than the π values for the same data (Figure 2-14, Figure 2-15, Figure 2-30, Figure 2-31, Table 2-3). This observation is most consistent with population expansion in Africans and a bottleneck or founder effect in Caucasians. The higher value of θ than π in Africans is an indication of the increased presence of low frequency variants relative to what is seen in Caucasians.

Table 2-2 Summary of population genetic parameters for Caucasians.

Section	Section Details	bp	θ*	π*	Tajimas D
Coding	Nonsynonymous	606,964	2.8	2.3	-0.66
Coding	All	789,502	4.2	3.8	-0.35
Coding	Synonymous	182,538	9.1	9.1	0.00
Intron	5' Intron (1-50)	245,646	6.5	6.9	0.23
Intron	5' Intron (7-50)	215,593	7.0	7.4	0.21
Intron	3' Intron (1-50)	245,384	6.5	6.6	0.06
Intron	3' Intron (7-50)	215,309	7.0	7.2	0.11
UTR	3' (1-50)	25,200	6.4	6.6	0.11
UTR	5' (1-50)	20,558	7.2	6.1	-0.55

*** Values of θ and π are 10^4. Numbers in parentheses represent the nucleotide positions used for analysis.**

Table 2-3 Summary of population genetic parameters for African Americans.

Section	Section Details	bp	θ*	π*	Tajimas D
Coding	Nonsynonymous	606,854	4.2	2.9	-1.12
Coding	All	789,502	6.6	5.1	-0.82
Coding	Synonymous	182,648	14.6	12.4	-0.55

Intron	5' Intron (1-50)	245,646	10.7	8.6	-0.72
Intron	5' Intron (7-50)	215,593	11.6	9.5	-0.66
Intron	3' Intron (1-50)	245,384	10.3	8.8	-0.54
Intron	3' Intron (7-50)	215,309	10.9	9.5	-0.47
UTR	3' (1-50)	25,200	8.4	7.6	-0.34
UTR	5' (1-50)	20,558	10.5	8.6	-0.65

* **Values of θ and π are 10^4. Numbers in parentheses represent the nucleotide positions used for analysis.**

Tajima’s D was calculated for all datasets and all populations combined for the 5’ and 3’ splice site sequences for the original and smoothed data (Figure 2-34 through Figure 2-37). These distributions were dominated by the polymorphism rates in the African Americans, since Africans have the highest polymorphism rates. As expected, D was uniformly negative for all hexamers (Figure 2-36, Figure 2-37). While the most negative values generally coincide with the regions where θ and π values are most reduced, the majority of Tajima’s D values are not below -2, which is an accepted cutoff for statistical significance. However, the general patterns observed, particularly for the 3’ splice site data, recapitulate the original patterns observed for θ and π, where D is most negative at hexamer start positions -1, -25, and -32. The pattern on the 5’ splice site side is less clear, although D appears reduced at position 1, as expected. The fact that θ is reduced more than π at locations where both are reduced is consistent with purifying selection having occurred at these locations.

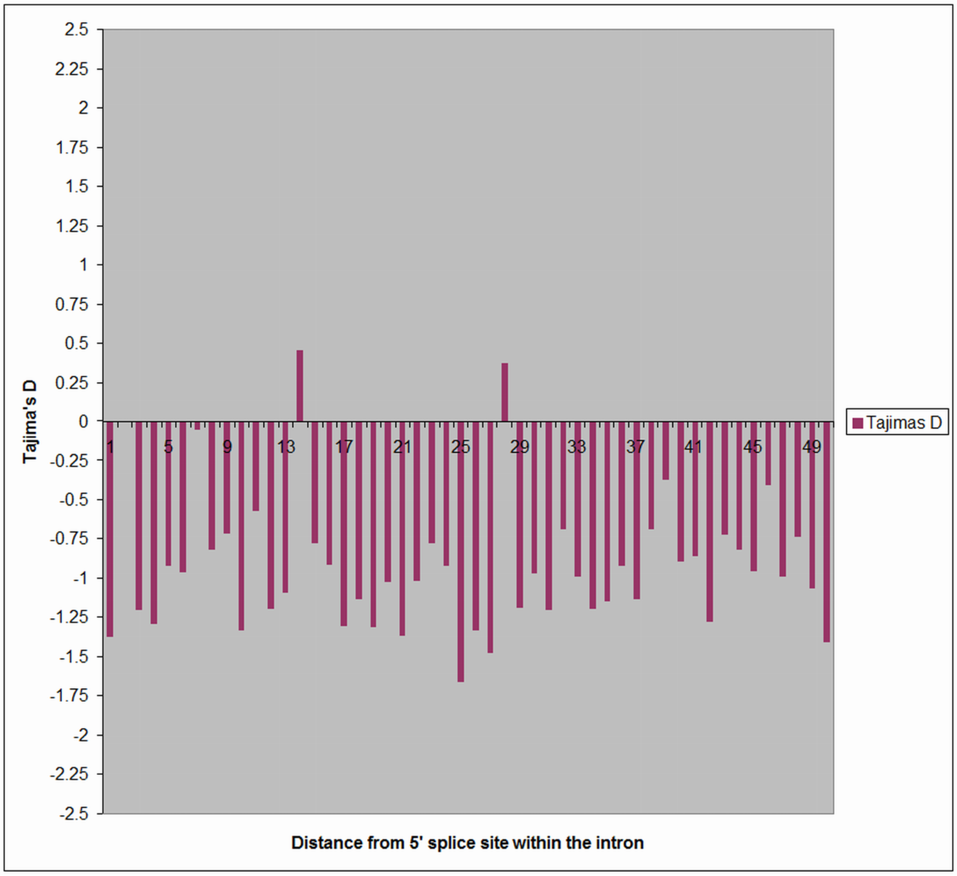

Figure 2-34 Distribution of Tajima's D values for the 5' splice sites using the combined datasets. Tajima's D values are shown for each nucleotide position.

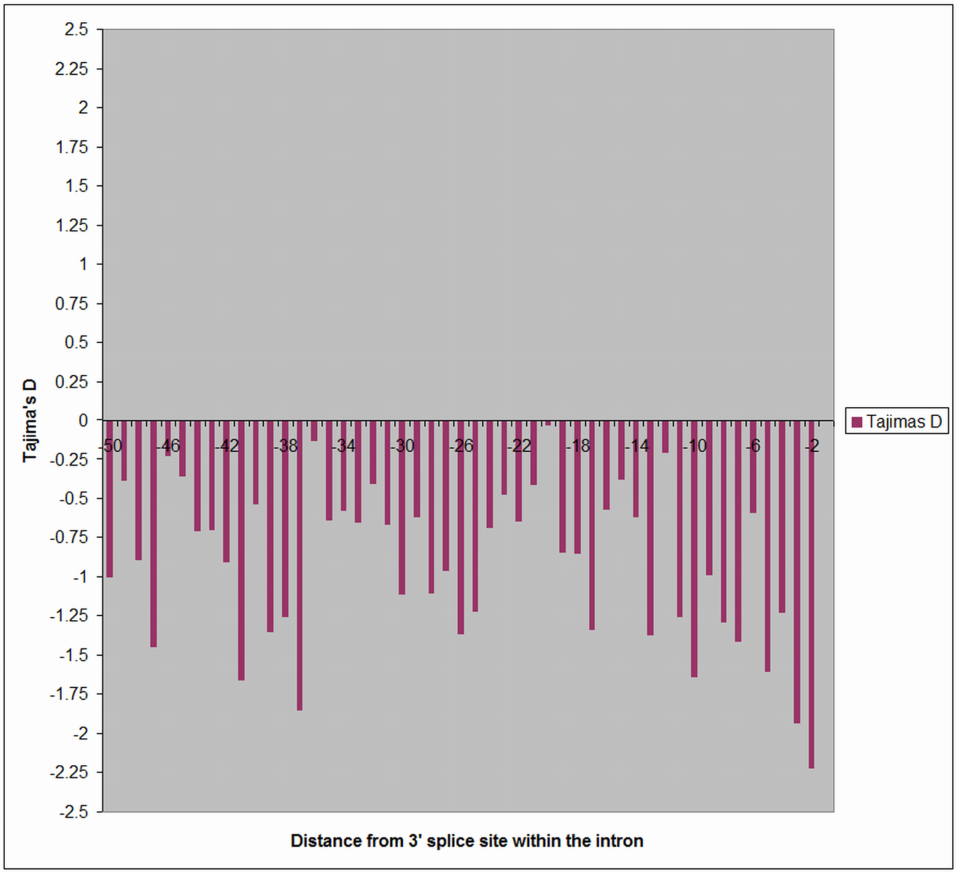

Figure 2-35 Distribution of Tajima's D values for the 3' splice sites using the combined datasets. Tajima's D values are shown for each nucleotide position.

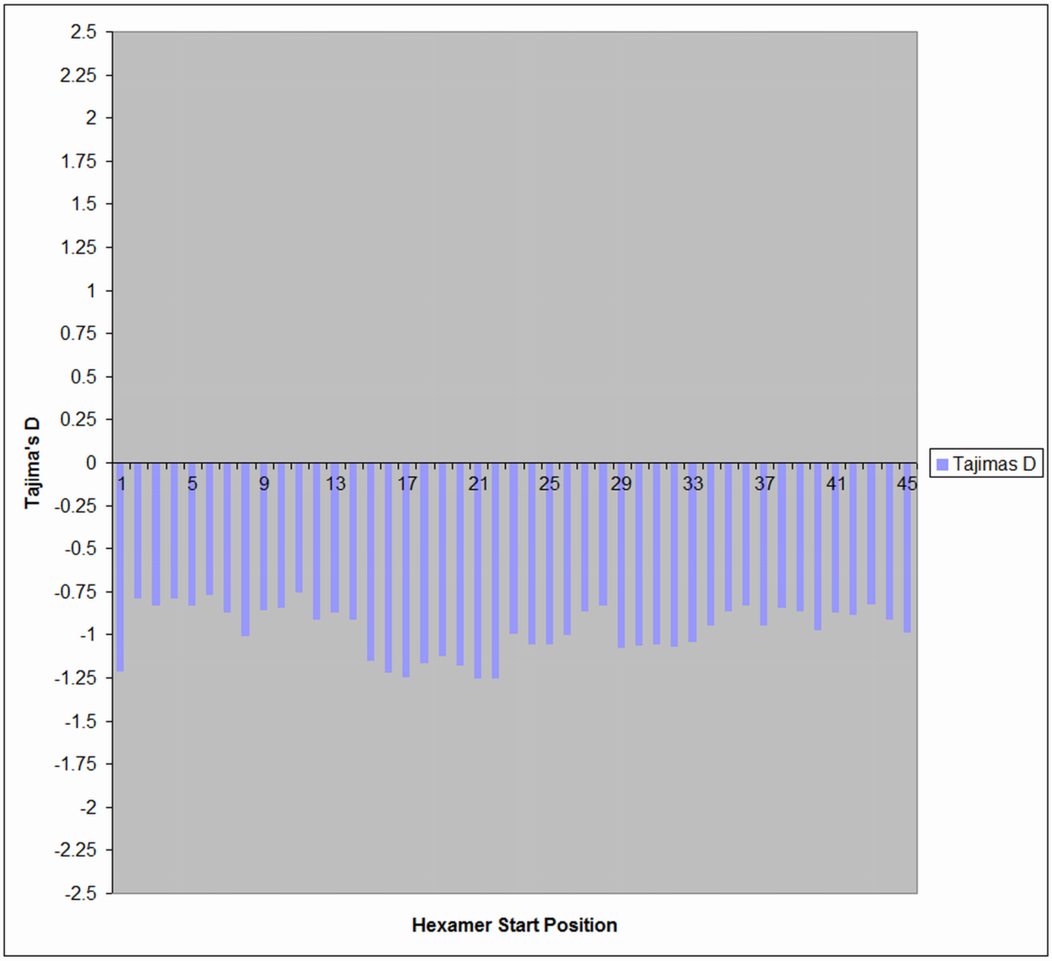

Figure 2-36 Distribution of Tajima's D values for hexamers in the 5' splice sites using the combined datasets. Tajima's D values are shown for each hexamer starting at the shown nucleotide positions.

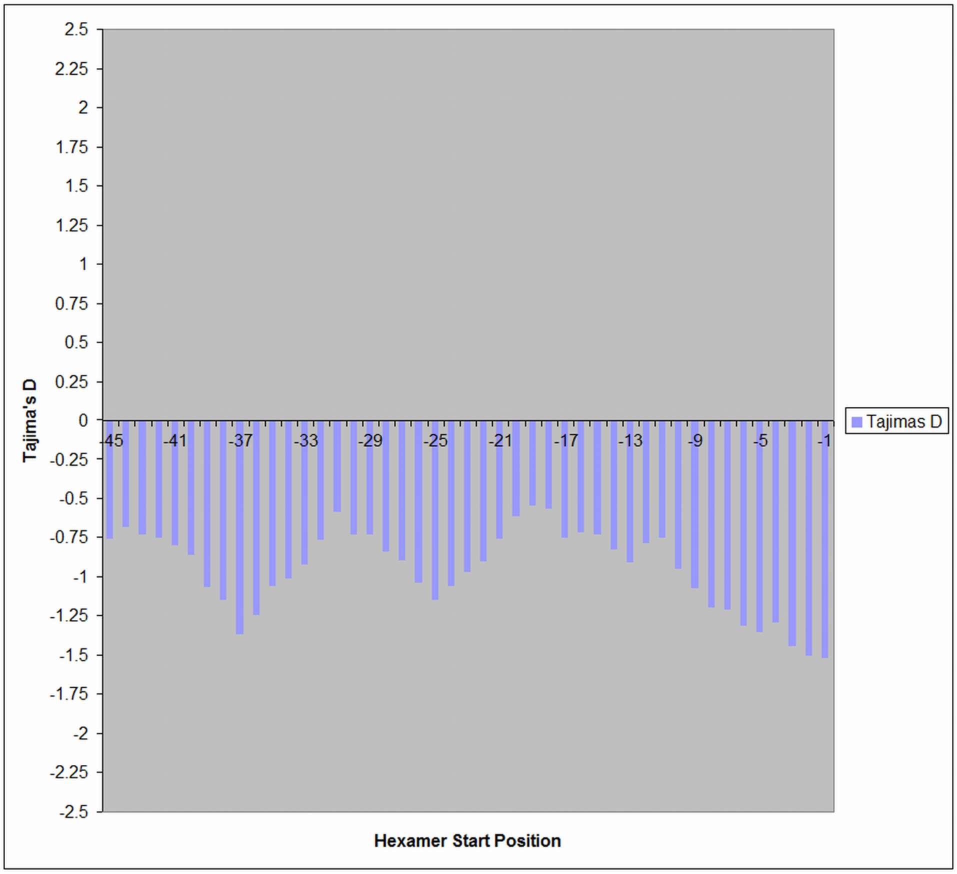

Figure 2-37 Distribution of Tajima's D values for hexamers in the 3' splice sites using the combined datasets. Tajima's D values are shown for each hexamer starting at the shown nucleotide positions.

2.4 Discussion

Using the intronic polymorphism data from three different databases, we have shown that human genetic variation can be used to identify regions of functional importance within non-coding regions. The intron motifs that are known to bind with various spliceosome elements - the 5' and 3' splice sites, branch site region, and polypyrimidine tract - all show reduced polymorphism levels that are unlikely to be observed within nonfunctional elements. It is commonly accepted that mutations of synonymous sites in coding regions are neutral, since a mutation at these sites will not change the amino acid sequence[56].

Therefore, we used the θ and π values observed within these positions to generate distributions of non-functional variation. Also, prior studies have shown that synonymous sites have the largest θ and π values of all coding, 5' UTR, and 3' UTR sites[17,18]. We note, however, there are a number of caveats to using this distribution for measuring statistical significance. Mutations at synonymous sites can: (1) stabilize mRNA secondary structure, which can beneficially prevent premature degradation or impede translation[56,57]; (2) alter regulatory splicing motifs such as exonic splicing enhancers and silencers[58]; and (3) change protein structure through the modification of translation rates due to codon usage bias[59]. For these reasons, distributions of θ and π values from synonymous sites might underestimate the true distribution of θ and π values for non-functional regions. However, this would result in an increase of false negatives and decrease of false positives in terms of inferring functional sites relative to using a true non-functional distribution. Therefore, if anything, our analyses would be conservative in terms of inferring functional sites based on reduced polymorphism.

While the 5' and 3' splice site are identified perfectly due to their precise location once the intron/exon boundaries have been recognized, the properties of other elements whose position are more variable, such as the branch site, will not always align to a specific location. Nonetheless, the signature of reduced polymorphism levels associated with functional motifs can still be observed. For example, the region which is most likely to contain the branch site, -23 through -32 (and possibly -35 through -44) from the 3' splice site, shows lower levels of polymorphism than other areas further from the intron/exon boundary. A more detailed investigation of the polymorphism properties of the branch site however will need to be conducted to confirm that the reduced polymorphism we observed at this location is indeed due to the branch site, and to further investigate the sequence properties from nucleotide -35 through -44. Because of the degeneracy of the branch site sequence and its variable positioning, a non-position-dependent measure will need to be used. The polypyrimidine tract's positioning on the other hand is more stable and the suppressed genetic variation from positions -6 through -15 can reliably be attributed to the presence and functional importance of the polypyrimidine tract.

Another finding was the prediction of a novel distance dependent localization for a functional region downstream of the 5' splice site, which appears generalized to most introns within genes. The reduced polymorphism levels constrained from nucleotide positions +21 through +30 downstream of the 5' splice site are consistent with observations from other studies that disease causing mutations can occur downstream of the 5' splice site in the +21 through +32 nucleotide range[60,61,62,63]. The cause for these diseases is the mutation of intronic splicing enhancers at these locations that decrease splicing efficiency and cause exon skipping. Intronic and exonic splicing enhancers and silencers (ISE, ISS, ESE, ESS) are motifs that can regulate splicing by promoting or inhibiting the retention of exons within genes[64]. These cis-acting elements are motifs located within both introns and exons that serve as binding regions for members of the SR family of proteins, which can recruit or inhibit the binding of different components of the spliceosome[64]. There are several examples of ISEs located downstream from exons that are specific to certain gene classes and intron positions[65]. The fact that this analysis included different gene classes and all their introns suggests that there may be a generalized distance dependent localization for functional motifs as other studies have suggested[66]. Future studies will need to characterize whether and which of the specific motifs within the +21 through +30 nucleotide range are responsible for the reduced polymorphism observed in this region.

While the genetic variation observed in the first six nucleotides of both the 3' and 5' splice sites was highly reduced, one notable observation was at position -4 of the 3' splice site. Prior studies have suggested that this position is not conserved within humans or across species[39]. However, our analyses did demonstrate reduction of genetic variation at this position, although only of modest statistical significance. This suggests that studies of genetic variation in humans may be able to detect regions of intermediate functional importance which may be missed through multi-species comparative methods. This trend, however, was not observed within every individual dataset. The Seattle SNP and PMT databases showed the higher variation consistent with a non-conserved position. This may be in part due to a functional requirement of that position within a specific class of genes such as DNA repair and cell cycle genes. The low variation observed in this position may also be due to a long distance correlation between the nucleotides observed in this position with other neighboring positions. Therefore, even though this position is generally not conserved, nucleotide sequences elsewhere in the gene may influence the nucleotide required for splicing in position -4.

As expected, we observed greater genetic variation in African Americans than Caucasians, and also more negative values of Tajima's D in African Americans. Although the location-specific patterns of variation we observed were consistent between the two groups, to some degree the patterns we observed were clearer in African Americans than Caucasians. This suggests that the negative selection we observed is not recent, predating the migration of European ancestors from Africa tens of thousands of years ago. Thus, African and African American populations may be most useful for identifying sites of historical selection and functional elements of the genome that are relevant to all populations.

2.5 Materials and Methods

2.5.1 Data

Sequence data were collected from three different sources. The UCSF Pharmacogenetics of Membrane Transporters (PMT) project contains sequence data for 45 human membrane transporter genes that were generated by sequencing 100 African Americans, 100 Caucasians, 30 Asians, 10 Hispanics, and 7 Pacific Islanders[67]. The Seattle SNP database is composed of sequence data for 290 genes involved in human inflammatory response obtained by sequencing 24 African Americans and 23 Caucasians[68]. The NIEHS SNP database is composed of sequence data for 386 genes involved in DNA repair and cell cycle pathways obtained by sequencing 90 individuals that were representative of the U.S.

population and include European-Americans, African-Americans, Mexican-Americans, Native-Americans, and Asian-Americans (NIEHS phase 1) in an undisclosed proportion. An additional set from the NIEHS database was also used that is composed of sequence data from 222 genes obtained by sequencing 95 individuals: 27 African-Americans, 22 Caucasians, 22 Mexican-Americans, and 24 Asian-Americans (NIEHS phase 2)[69].

The amount of intron sequence differed among databases. The PMT amplified each exon with a minimum of 35 flanking 5' and 3' intronic basepairs[67]. The Seattle SNP genes were sequenced for both exons and full introns and about 2,500 base pairs upstream and about 1,500 base pairs downstream of the genes[68]. The NIEHS project fully sequenced genes (including introns) that were <30 kb, whereas for larger genes all coding and conserved noncoding sequence, as well as 20% of the remaining intronic sequences were targeted[69].

2.5.2 Population Genetic Parameters

When a nucleotide position within a gene is mutated and causes an amino acid change, devastating effects on protein function can take place, particularly if that residue was required for binding or catalysis. For this reason, these positions will tend to have reduced variation, or decreased polymorphism levels, in the human population. Using this same principle, other sites with decreased genetic variation have been suggested to be under negative selection and therefore are likely indicators of functionally important regions within the genome. Consequently, measurement of nucleotide diversity, which provides a measure of genetic variation, is commonly employed to infer functional sites.

For this study, nucleotide diversity was measured using the two population genetic parameters θ and π (e.g. see Hartl and Clark[70]). θ represents the standardized proportion of segregating sites in a sequence and π is the average proportion of nucleotide differences per site between all pairs of chromosomes in the sample. The formulas for θ and π are given by:

$$\hat{\theta} = \frac{S}{a_1}$$

and $$\hat{\Pi} = \frac{\sum_{i=1}^{L_{seq}} X_i(n - X_i)}{\sum_{i=1}^{L_{seq}} \binom{n}{2}}$$

where $$S = \frac{x_{poly}}{L_{seq}} \qquad a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

and L_{seq} = sequence length, x_{poly} = number of polymorphic sites, n = number of chromosomes, and X_i = number of chromosomes carrying the variant allele at position i of *seq*.

These measurements were used because they are normalized for both sequence length and sample size. Both are estimates of the population genetic parameter $4N_e\mu$, where N_e is the effective population size and μ is the per site mutation rate per generation. Under a model of neutral evolution both θ and π should be equal. Significant differences between both θ and π can indicate natural selection or population expansion or contraction. Both θ and π were measured using the available sequence data from each database for all individuals and ethnicities. Because θ and π depend on population history, we also examined these distributions separately in the two largest ethnic groups comprising this sample, namely African Americans and Caucasians.

One commonly employed method to test if a given sequence is evolving neutrally or under some type of selection is the statistic known as Tajima's D[71]. The premise of this statistic is the expected difference between π and θ under varying conditions. Under neutral evolution of a population of constant size, D is expected to be 0. Under purifying selection, negative values of D are expected, although recent population expansion can also create negative values of D. Positive values of D occur under balancing selection or recent population decrease. The formula for Tajima's D is given by:

$$D = \frac{\hat{\Pi} - \hat{\theta}}{\sqrt{\left[c_1\hat{\theta}/k + ea_1\hat{\theta}(a_1\hat{\theta} - 1/k)\right]}} \text{ where}$$

$$e = \frac{c_2}{a_1^2 + a_2} \qquad c_1 = b_1 - \frac{1}{a_1} \qquad c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}$$

$$b_1 = \frac{n+1}{3(n-1)} \qquad b_2 = \frac{2(n^2+n+3)}{9n(n-1)} \qquad a_2 = \sum_{i=1}^{n-1}\frac{1}{i^2} \qquad k = L_{seq}$$

2.5.3 Comparison of θ and π Across Genomic Regions

To compare mean θ and π values between different gene regions, we employed t-tests. However, θ and π values are influenced by the racial/ethnic composition of each individual study, and potentially by the class of gene. Therefore, to control for these potential differences, we used a matched pair t-test, where we compared θ or π values between regions within each gene (e.g. exons versus introns), and then evaluated the mean value of the derived t-statistic. The standard errors of mean θ and π values reported in Table 1 were calculated across all genes.

2.5.4 Statistical Significance of Population Genetic Parameters

While the approach above is appropriate for nucleotide sequences of sufficient length, it is inadequate for studying single nucleotides because the large majority of θ and π values will be 0. We therefore took the following approach. For nucleotides at a specific sequence location, we concatenated them into a single sequence, and calculated the θ and π values for that created sequence. Then to be able to distinguish between nucleotide positions in the genome that are neutrally evolving and those undergoing selection, a reference empirical distribution of the θ and π values associated with nucleotides from non-functional regions created in a similar fashion is required. Any measured values can then be compared to this non-functional distribution.

Distributions were generated for each of the four data sets by taking 10,000 random samples of synonymous sites and calculating θ and π values for each sample. Only four-fold degenerate (synonymous) sites from each database were used during the random sampling and separate distributions were made for θ and π. Theoretically, both θ and π normalize for sequence length but for small lengths it can be difficult to differentiate between a functional and non-functional sequence because variation is expected to be low even in a large sample. For example, the probability of seeing a sequence of length 50 with no SNPs of frequency 1% or greater is 60.5%, whereas for a sequence of length 500 the same probability drops to .66%. For this reason, length based distributions of θ and π were generated that corresponded to the length (L) of the query sequence that was being tested for functionality – each of the 10,000 θ and π values were calculated from L random data points. For analyses focusing on a single ethnic group, such as African Americans or Caucasians, distributions were generated for each of these groups separately.

2.5.5 Selection of Nucleotide Regions

A number of factors influenced the choice of nucleotide regions selected for analyses. Due to the limited length of introns sequenced within the PMT database and the known location of some functional motifs within introns, the 50 basepairs flanking exons from all introns were measured for genetic variation. The same numbering nomenclature shown in Figure 2-1 was used. The 5' and 3' splice sites, also known as the donor and acceptor sites, respectively, refer to the intron/exon boundaries located at the 5' and 3' parts of an intron. Numbering is relative to the location of the splice site. For instance, position +2 at the 5' splice site is two bases into the intron whereas position +2 at the 3' splice site is two bases into the exon. Genetic variation was measured from positions +1 through + 50 of the 5' splice site and -50 through -1 of the 3' splice site at every individual position. For example, the genetic variation measured at position -7 was taken from every intron within every gene from every database at position -7 relative to the 3' splice site. The results of measuring the nucleotide diversity within these predefined positions relative to the splice sites implies that the known intronic features that are located at the 5' and 3' splice site boundaries will be well aligned. On the other hand, features such as the branch site that occur further within the intron and whose location is not confined within a narrow range may be blurred due to using this distance dependent intron alignment. Intron/exon boundaries were previously defined for every gene within each database, thus the same boundaries were used in our analyses.

2.5.6 Selection of Sliding Windows of Length 6

Analysis of single sites within introns produced large variability. Therefore, to obtain a clearer picture of regional variation within introns, we also examined sliding windows of 6 adjacent nucleotides. The choice of 6 nucleotides was based on a compromise between the need for smoothing and the potential loss of site-specific variation. Thus, this approach will detect reductions in variation that are regional as opposed to single site-specific, although signals may also emerge for single sites for hexamers that overlap that single site provided the single-site reduction is large. Sliding windows of length 3 and 4 were also generated and were found to produce nearly identical results to those of length 6 (data not shown).

3 Characterizing the Branch Site and Intronic Splicing Enhancers

While we have been successful in using genetic variation to characterize known functional non-coding elements, we have discovered additional regions of unknown function that we believe to be functional. We hypothesized that the branch site is the cause for at least one of these regions, while the additional regions are localization sites of intronic splicing enhancers. In the following chapter, we present the additional studies conducted to further characterize the cause of the reduced polymorphism levels observed in those regions. To investigate this matter, we implemented computational methods that included a Hidden Markov model trained to predict branch sites, as well as a novel motif finder based on identifying motifs of reduced genetic variation.

3.1 Abstract

In a previous study, we found that measurement of genetic variation within non-coding genomic sequences can be used to recognize functional genetic elements. While several regions of reduced genetic variation corresponded precisely with known functional motifs important for splicing, the remaining areas did not match any well-characterized regions of functional significance. In this study, we further examine the properties of these latter areas in more detail, in order to characterize the reasons for their reduced genetic variation. To aid in this task, we developed a novel motif-finding algorithm based on finding motifs with reduced genetic variation. We also developed a Hidden Markov model and used

it in conjunction with an already established branch site prediction algorithm to predict the precise location of branch sites upstream of the 3' splice site. Using these computational tools, we confirmed that one of the previously predicted functional regions upstream of the 3' splice, from positions -23 to -27, corresponded to a preferred localization site for branch sites. Further, we conclude that two nucleotides within the branch site are important for splicing. Using our motif-finding algorithm, we reinforced our original prediction that the area from nucleotides -35 to -39 on the 3' splice site and +21 to +30 on the 5' splice site correspond to preferred localization sites for intronic splicing enhancers.

3.2 Introduction

A major goal in human genetics research is to identify disease or trait-causing alleles in the human genomes. Historically, much of the emphasis has focused on studying coding, as opposed to non-coding variants, when searching for disease-causing alleles. While it is well established that reduced population genetic variation at a nucleotide site is a good indicator of a site's functional significance within coding regions, we previously showed that similar analyses are able to distinguish functional from non-functional elements within non-coding regions, most notably within introns (2). In that analysis, we characterized nucleotides based on their precise position, that is, distance from the exon-intron boundary. A potential limitation of that analysis was that functional motifs within the intron were not characterized in detail; if such motifs vary in their physical locations, our power to detect and characterize functional sites within them would be reduced.

One such example is the branch site, an important motif that aids in intron identification, spliceosome formation, and lariat formation during mRNA splicing. The localization of the branch site is highly variable, but confined to 15 to 50 basepairs upstream of the 3’ splice site, since the spliceosomal element that binds to the branch site must first be recruited by elements that recognize the 3’ splice site[30]. In our previous analyses, polymorphism levels were measured for each position of the 3’ splice site region as a function of its distance from the intron/exon boundary; therefore, given the variable positioning of the branch site, the measurement of the genetic variation within the intron ends was only able to elucidate what might be generalized preferences for the branch site distance from the 3’ intron-exon boundary.

In our previous analyses, we showed that polymorphism levels for the first 6 nucleotides of the 5’ splice site were significantly reduced, confirming their functional importance in proper exon splicing (2.3.2 and 2.3.3). However, we also identified one additional region in the 5’ portion of introns that appeared to have reduced genetic variation, namely from nucleotide position +24 through +30. Thus, this region may harbor sequences of functional significance within this general location across introns.

Similarly, in the region of the 3’ splice site, we identified a longer track of nucleotides with reduced polymorphism, from position -1 to approximately position -15. The first six nucleotides likely reflect the splice site recognition signal, similar to the one we observed on the 5' side of the intron. The remaining positions include the polypyrimidine tract (PPT), another known motif important for proper intron splicing. The PPT, a stretch of ten or more nucleotides with a preference for pyrimidines that serves as a binding site for the U2AF snRNP, is one of the many factors required for intron splicing[30]. However, we also identified two additional regions with reduced polymorphism, one spanning sites -23 to -27, and another spanning sites -35 to -39. We conjectured that the first of these represents a preferential location for the branch site, while the latter may represent a second preferential location for the branch site or the preferential location of another functional motif.

It is the focus of the current chapter to examine these intronic regions for specific sequence motifs that represent functional elements, and to determine whether these motifs can explain the regions of reduced polymorphism that we previously identified. To do so, for known motifs (such as the branch site) we trained a Hidden Markov model to identify the most likely location of that motif within a given sequence. For previously undefined motifs, we employed a novel motif finder, based on characterization of nucleotide sequences in a specific region that demonstrate reduced polymorphism. Specifically, we focused our search on the branch site, polypyriminde tract, and splicing enhancer motifs as possible explanations for regions of reduced genetic variation within the 5' and 3' regions of introns.

3.2.1 The Branch Site

The branch site is highly degenerate yet faithfully recognized by the spliceosome as the site of lariat formation by base pairing with the U2 snRNP. U2 contains a highly conserved non-variant motif (GUAGUA) within both humans and yeast[72]. This motif is not only able to bind to the invariant branch site UACUA**A**C in yeast (the bolded branch point site bulges out during binding) but also to the more degenerate YUR**A**Y (Y represents any pyrimidine, R represents any purine) element found within mammalian pre-mRNAs[72]. Furthermore, the more similar the human branch site is to the yeast branch site motif, the more efficiently splicing occurs[72]. The lack of a perfect complementary base pairing mechanism between U2 and the branch site in humans is possible because of the requirement of several other splicing factors that aid in spliceosome recruitment. Therefore, to obtain more detailed insight into the functionally important positions within the branch site, given its overall degeneracy, the branch site motif needs to be predicted from verified branch sites, and its genetic variation measured.

3.2.2 Previous Work

The identification of the branch site has been widely investigated. One of the earliest studies that used computational methods for branch site prediction used a position specific scoring matrix from the yeast consensus sequence to identify high scoring sequences in different organisms[73]. Using those high scoring sequences, a new consensus sequence was generated and used to identify other potential branch sites, selecting the sequence closest to the splice site if several were predicted. Every predicted sequence was then aligned and a new branch site profile for each organism was reported. Position specific scoring matrices (PSSMs) have found their way into other branch site prediction studies, such as one by Lim et al. where the branch site consensus sequence for a PSSM was generated from potential branch sites identified using a Gibbs sampling algorithm[74]. Although such methods are sufficient to generate a profile for branch sites, their accuracy is questionable since the branch site is highly degenerate and its position is highly variable across introns and genes, making correct predictions with a PSSM difficult[75].

A more recent study continued the use of a PSSM to identify high scoring branch site sequences, but improved the method by looking for branch sites within 6 nucleotides of predicted PPTs (refered to as PSSM-polyY)[40]. Results from that study showed that 80% of the positive controls were correctly called using their method; however, in an extended dataset of 46 controls[24,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55] only 68% were correctly called, due to inclusion of sequences with weak PPTs.

3.2.3 The Polypyrimidine Tract

As noted above, another important motif located within the intron is the polypyrimidine tract (PPT) [30]. In the current study, we not only document the reduced variation generally observed for this motif, but also provide a more detailed analysis characterizing the variation in terms of transitions versus transversions of pyrimidine and purine sites. Specifically, we evaluate transition bias, which is the ratio of the rate of transitions (Y→Y, R→R) to the rate of transversions (Y→R, R→Y), defined as TI:TV[76]. Assuming equal substitution rates at all nucleotides, the ratio should be equal to .5 because there are twice as many possibilities for a transversion as a transition; however, it has long been recognized that transitions are more common than transversions[76]. Example values of the TI:TV ratio are 2.3[77], 2.44[14] and 3.02[78] within coding regions and 1.99[78] within noncoding regions. Explanations for this observation include: (1) transitions within the 3rd codon position cause amino acid substitutions only 3% of the time compared to 41% of the time with transversions, (2) transitions at the 1st and 2nd codon positions tend to create conservative amino acid substitutions that maintain the chemical properties of amino acids, and (3) the free energies of base pairing in transition mutations, G-T and A-C, are more favorable than in transversion mutations[76].

3.2.4 Splicing Enhancers

Results from our previous analysis suggest the possible existence of intronic splicing enhancers within both the 5’ and 3’ splice site regions (2.4). Given the lack of concrete information about functional motifs underlying splicing enhancers located within these regions, we instead present an algorithmic search for motifs that is based on reduced genetic variation. We then compare the results of this algorithmic search with formerly defined motif characteristics of splicing enhancers, and also determine whether the location of these motifs in the 5’ and 3’ splice site regions overlap with the regions of reduced polymorphism that we previously observed.

3.3 Results

3.3.1 Polypyrimidine Tract

The genetic variation of the polypyrimidine tract, as expected, is lower than the intronic controls for both θ and π (Table 3-1 and Table 3-2). This same pattern was observed within the individual datasets of NIEHS P1, NIEHS P2, PMT, and Seattle SNP (Table 3-3 through Table 3-6). When nucleotide positions within this region are separated into their classes, purines show higher genetic variation whereas pyrimidines show lower variation than every control site. Similar trends were observed within the individual datasets where purine mutation rate is high and pyrimidine mutation rate is low (Table 3-3 through Table 3-6). Given the difference in levels of genetic variation between purines and pyrimidines and to determine if certain types of mutations were favored over others, both purines and pyrimidines are further classified into groups depending on the basis of the nucleotide change. TI:TV ratios are calculated for both pyrimidines and purines using both θ and π results from Table 3-1.

Table 3-1 Summary of π and θ values for the different nucleotide classes within the polypyrimidine tract and three control regions. Values are x10^4.

				π					
Region	**Overall**	**R^a**	**R_{TI}^b**	**R_{TV}^c**	**$R_{TI:TV}^d$**	**Y^e**	**Y_{TI}^f**	**Y_{TV}^g**	**$Y_{TI:TV}^h$**
PolyY Tract	6.5	10.4	6.0	3.2	1.88	5.5	3.2	1.5	2.13
Control 3' SS (27-34)	8.1	9.5	6.4	2.5	2.56	6.9	4.7	1.5	3.13
Control 5' SS (7-14)	9.0	9.9	6.2	2.5	2.48	8.3	5.8	1.8	3.22
Control 5' SS (27-34)	8.7	9.0	6.1	2.3	2.65	8.2	5.5	2.2	2.50

				θ					
Region	**Overall**	**R^a**	**R_{TI}^b**	**R_{TV}^c**	**$R_{TI:TV}^d$**	**Y^e**	**Y_{TI}^f**	**Y_{TV}^g**	**$Y_{TI:TV}^h$**
PolyY Tract	10.1	14.5	8.2	4.8	1.71	8.9	5.5	2.4	2.29
Control 3' SS (27-34)	10.8	12.3	8.6	2.8	3.07	9.6	6.4	2.5	2.56
Control 5' SS (7-14)	12.0	12.9	8.3	3.5	2.37	11.2	7.7	2.7	2.85
Control 5' SS (27-34)	12.0	13.3	8.5	4.1	2.07	10.8	7.0	2.9	2.41

[a]R: Purines, [b]R_{TI}: Purine Transitions, [c]R_{TV}: Purine Transversions, [d]$R_{TI:TV}$: Purine transition/transversion ratio, [e]Y: Pyrimidines, [f]Y_{TI}: Pyrimidine Transitions, [g]Y_{TV}: Pyrimidine Transversions, [h]$Y_{TI:TV}$: Pyrimidine transition/transversion ratio

Table 3-2 Summary of π and θ percentiles for the different nucleotide classes within the polypyrimidine tract and three control regions.

				π Percentiles			
Sites	**Overall**	**R^a**	**R_{TI}^b**	**R_{TV}^c**	**Y^e**	**Y_{TI}^f**	**Y_{TV}^g**
PolyY Tract	0	.203	.035	.392	0	0	.002
Control 3' SS (27-34)	.098	.019	.015	.055	.224	.401	.01
Control 5' SS (7-14)	.543	.041	.006	.056	.836	.92	.056
Control 5' SS (27-34)	.359	.004	.004	.023	.811	.827	.234
				θ Percentiles			
Region	**Overall**	**R^a**	**R_{TI}^b**	**R_{TV}^c**	**Y^e**	**Y_{TI}^f**	**Y_{TV}^g**
PolyY Tract	0	.233	.003	.88	0	0	.003
Control 3' SS (27-34)	.001	0	0	.006	.001	0	.027
Control 5' SS (7-14)	.166	.001	0	.171	.155	.043	.089
Control 5' SS (27-34)	.161	.006	0	.631	.072	.003	.212

Table 3-3 Summary of π and θ values for the different nucleotide classes within the polypyrimidine tract and four control regions for the NIEHS Phase 1 data. Values are x10^4.

					π				
Region	**Overall**	**R**[a]	**R_{TI}**[b]	**R_{TV}**[c]	**$R_{TI:TV}$**[d]	**Y**[e]	**Y_{TI}**[f]	**Y_{TV}**[g]	**$Y_{TI:TV}$**[h]
PolyY Tract	5.5	8.5	5.4	1.9	2.84	4.7	2.9	1.0	2.90
Control 3' SS (27-34)	7.0	9.0	6.1	2.4	2.54	5.5	4.0	1.3	3.08
Control 5' SS (7-14)	7.3	9.4	6.8	1.8	3.78	5.4	4.2	0.7	6.00
Control 5' SS (27-34)	7.2	7.5	4.8	2.1	2.29	6.9	4.5	1.9	2.37

					θ				
Region	**Overall**	**R**[a]	**R_{TI}**[b]	**R_{TV}**[c]	**$R_{TI:TV}$**[d]	**Y**[e]	**Y_{TI}**[f]	**Y_{TV}**[g]	**$Y_{TI:TV}$**[h]
PolyY Tract	8.8	12.7	6.5	4.1	1.59	7.7	5.1	2.0	2.55
Control 3' SS (27-34)	9.8	11.6	7.8	2.6	3.00	8.3	5.7	2.4	2.38
Control 5' SS (7-14)	10.4	12.1	8.4	2.7	3.11	8.9	6.4	1.8	3.56
Control 5' SS (27-34)	11.5	13.2	8.0	4.3	1.86	9.8	6.1	2.7	2.26

[a] R: Purines, [b] R_{TI}: Purine Transitions, [c] R_{TV}: Purine Transversions, [d] $R_{TI:TV}$: Purine transition/transversion ratio, [e] Y: Pyrimidines, [f] Y_{TI}: Pyrimidine Transitions, [g] Y_{TV}: Pyrimidine Transversions, [h] $Y_{TI:TV}$: Pyrimidine transition/transversion ratio

Table 3-4 Summary of π and θ values for the different nucleotide classes within the polypyrimidine tract and four control regions for the NIEHS Phase 2 data. Values are x10^4.

					π				
Region	**Overall**	**R**[a]	**R_{TI}**[b]	**R_{TV}**[c]	**$R_{TI:TV}$**[d]	**Y**[e]	**Y_{TI}**[f]	**Y_{TV}**[g]	**$Y_{TI:TV}$**[h]
PolyY Tract	6.9	11.9	6.4	4.0	1.60	5.6	2.9	1.8	1.61
Control 3' SS (27-34)	8.4	9.8	6.8	2.8	2.43	7.3	4.8	1.7	2.82
Control 5' SS (7-14)	10.0	8.8	5.0	2.7	1.85	11.1	8.1	2.5	3.24
Control 5' SS (27-34)	8.2	7.5	6.5	1.0	6.50	8.9	4.5	2.9	1.55

					θ				
Region	**Overall**	**R**[a]	**R_{TI}**[b]	**R_{TV}**[c]	**$R_{TI:TV}$**[d]	**Y**[e]	**Y_{TI}**[f]	**Y_{TV}**[g]	**$Y_{TI:TV}$**[h]
PolyY Tract	11.3	13.7	8.1	5.1	1.59	10.7	6.4	2.9	2.21
Control 3' SS (27-34)	11.6	13.1	9.5	2.9	3.28	10.5	6.7	2.9	2.31
Control 5' SS (7-14)	14.1	14.3	8.7	4.2	2.07	14.0	9.0	4.2	2.14
Control 5' SS (27-34)	12.1	11.9	8.3	3.4	2.44	12.3	7.6	3.2	2.38

[a] R: Purines, [b] R_{TI}: Purine Transitions, [c] R_{TV}: Purine Transversions, [d] $R_{TI:TV}$: Purine transition/transversion ratio, [e] Y: Pyrimidines, [f] Y_{TI}: Pyrimidine Transitions, [g] Y_{TV}: Pyrimidine Transversions, [h] $Y_{TI:TV}$: Pyrimidine transition/transversion ratio

Table 3-5 Summary of π and θ values for the different nucleotide classes within the polypyrimidine tract and four control regions for the Seattle SNP data. Values are x10^4.

π

Region	Overall	R^a	R_{TI}^b	R_{TV}^c	$R_{TI:TV}^d$	Y^e	Y_{TI}^f	Y_{TV}^g	$Y_{TI:TV}^h$
PolyY Tract	8.3	12.3	7.8	3.8	2.05	7.3	4.5	2.2	2.05
Control 3' SS (27-34)	8.6	9.4	5.7	2.8	2.04	8.0	6.0	1.6	3.75
Control 5' SS (7-14)	9.6	10.0	6.3	3.6	1.75	9.3	6.4	2.8	2.29
Control 5' SS (27-34)	11.1	13.1	8.1	3.9	2.08	9.1	7.2	1.9	3.79

θ

Region	Overall	R^a	R_{TI}^b	R_{TV}^c	$R_{TI:TV}^d$	Y^e	Y_{TI}^f	Y_{TV}^g	$Y_{TI:TV}^h$
PolyY Tract	10.3	16.5	10.6	4.2	2.52	8.7	5.1	2.7	1.89
Control 3' SS (27-34)	11.0	11.8	8.1	2.9	2.79	10.4	7.3	2.5	2.92
Control 5' SS (7-14)	10.9	11.2	7.4	3.8	1.95	10.7	7.4	2.9	2.55
Control 5' SS (27-34)	12.7	15.1	9.4	4.8	1.96	10.4	7.6	2.5	3.04

[a]R: Purines, [b]R_{TI}: Purine Transitions, [c]R_{TV}: Purine Transversions, [d]$R_{TI:TV}$: Purine transition/transversion ratio, [e]Y: Pyrimidines, [f]Y_{TI}: Pyrimidine Transitions, [g]Y_{TV}: Pyrimidine Transversions, [h]$Y_{TI:TV}$: Pyrimidine transition/transversion ratio

Table 3-6 Summary of π and θ values for the different nucleotide classes within the polypyrimidine tract and four control regions for the PMT data. Values are x10^4.

π

Region	Overall	R^a	R_{TI}^b	R_{TV}^c	$R_{TI:TV}^d$	Y^e	Y_{TI}^f	Y_{TV}^g	$Y_{TI:TV}^h$
PolyY Tract	3.8	10.4	2.1	5.8	0.36	2.2	1.2	1.0	1.20
Control 3' SS (27-34)	10.9	10.8	8.8	0.4	22.00	9.6	4.5	1.3	3.46
Control 5' SS (7-14)	14.3	17.2	7.6	2.5	3.04	12.1	5.1	1.9	2.68
Control 5' SS (27-34)	10.2	7.8	4.8	1.3	3.69	11.2	8.9	2.3	3.87

θ

Region	Overall	R^a	R_{TI}^b	R_{TV}^c	$R_{TI:TV}^d$	Y^e	Y_{TI}^f	Y_{TV}^g	$Y_{TI:TV}^h$
PolyY Tract	12.8	20.9	9.7	9.7	1.00	10.4	6.5	2.9	2.24
Control 3' SS (27-34)	13.5	16.7	13.4	2.7	4.96	11.0	7.3	1.6	4.56
Control 5' SS (7-14)	17.4	18.9	10.0	4.1	2.44	16.1	11.1	2.2	5.05
Control 5' SS (27-34)	12.5	10.9	7.9	1.8	4.39	14.0	10.1	3.9	2.59

[a]R: Purines, [b]R_{TI}: Purine Transitions, [c]R_{TV}: Purine Transversions, [d]$R_{TI:TV}$: Purine transition/transversion ratio, [e]Y: Pyrimidines, [f]Y_{TI}: Pyrimidine Transitions, [g]Y_{TV}: Pyrimidine Transversions, [h]$Y_{TI:TV}$: Pyrimidine transition/transversion ratio

Both purine and pyrimidine TI:TV ratios are lower than any of the intronic controls suggesting a modified mutational behavior within the polypyrimidine tract, which can be attributed to an increase in transversions or a decrease in transitions. The decrease in TI:TV ratio for purines in the PPT can be principally attributed to larger increase of transversions versus transitions in the PPT compared to the control regions (Table 3-1). On the other hand, the decrease in TI:TV ratio for pyrimidines is due to the relative decrease in pyrimidine transitions in the PPT compared to the control segments. These data suggest the PPT's pyrimidines are under negative selection (i.e., new mutations within pyrimidines, either transitions or transversions, are removed) while the purines are under positive selection (i.e., new mutations within purines, specifically transversions, are favored).

3.3.2 Branch Site

Our method of characterizing the genetic variation of nucleotides within introns based on their precise position, that is, distance from the exon-intron boundary, can have some limitations. For instance, our power to detect and characterize functional sites within introns can be reduced if their physical locations are variable. Given the variable positioning of the branch site, our previously described analyses of the 3' splice site can only elucidate what might be generalized preferences for the branch site distance from the 3' intron-exon boundary. Therefore, to not only characterize regions -23 to -27 and -35 to -39, but also obtain more detailed insight into the functionally important positions within the branch site given its overall degeneracy, we trained a Hidden Markov model to identify its most likely location within a given sequence.

We then used this HMM jointly with the PSSM-polyY approach (3.5.3) to identify putative branch sites and then characterize the polymorphism level for the 5 nucleotides within the branch site pentamer (5'-YUR**A**Y-3'). Results are given in Figure 3-1 and Figure 3-2. It is clear that the adenosine (A) site is most reduced in genetic diversity within the branch site sequence (Figure 3-1 and Figure 3-2). This finding is consistent with the fact that the A site is highly conserved within human branch sites and directly participates in lariat formation and intron splicing[30]. The only other position that shows reduced polymorphism values is the U site, which is located 2 bp upstream of the conserved adenosine site.

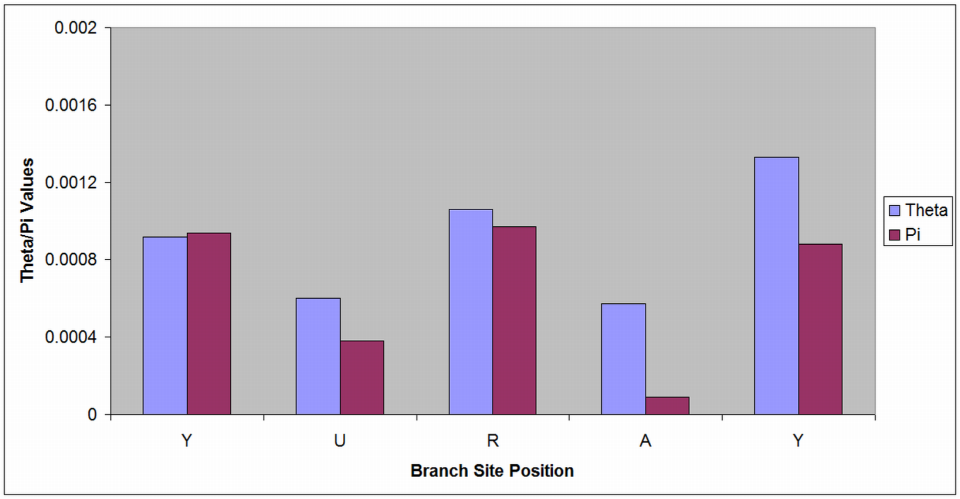

Figure 3-1 Distribution of polymorphism for the branch site predicted using the joint PSSM-polyY and HMM prediction method on the combined datasets. θ and π values are shown for each position of the branch site motif.

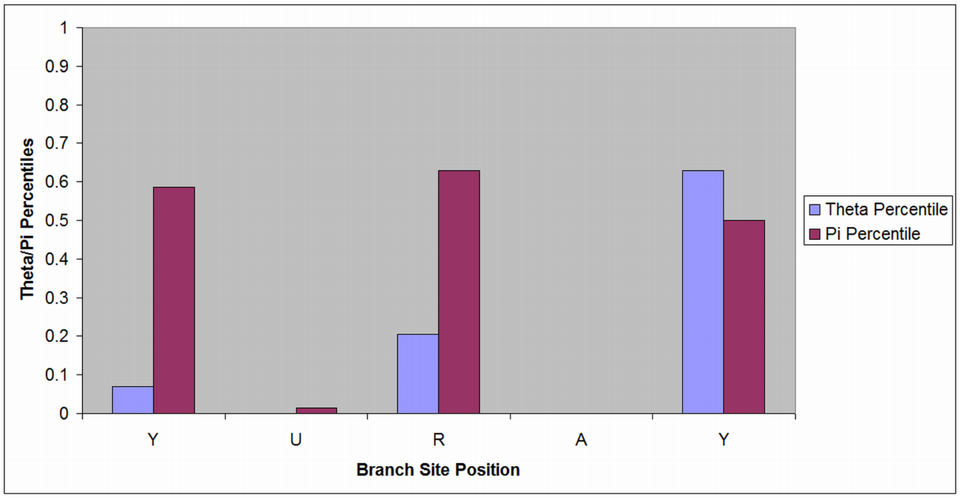

Figure 3-2 Distribution of polymorphism for the branch site predicted using the joint PSSM-polyY and HMM prediction method on the combined datasets. Percentiles are shown for each position of the branch site motif.

We then examined the frequency distribution for the predicted location of the conserved adenosine residue within the branch site from the results of our branch site prediction procedure. The distribution showed a peak at position -25, with a range of -22 to -26 (Figure 3-3). We note that this places the most likely location of the proximal conserved branch site uracil between positions -24 to -28. Therefore, in terms of reduced sequence variation, we would expect that the greatest reduction would occur at the average positions of the conserved adenosine and uracil which are positions -23 to -27. Examination of Figure 2-20 and Figure 2-21 from chapter 2 (2.3.3) shows that positions -23 to -27 correspond precisely to a peak region of reduced polymorphism that we first detected. As another test, we recalculated θ and π distributions for hexamers for the 3' introns, but in this case only included the introns for which we had a branch site prediction. In addition, we removed the predicted branch sites from this analysis, to determine the extent to which the predicted branch sites could explain the previously observed localized pattern of reduction in genetic diversity. The results are provided in Figure 3-4 and Figure 3-5. Compared to Figure 2-20 and Figure 2-21 (2.3.3), there is a clear attenuation of the previously seen reduction of diversity at hexamer start sites -23 to -27, which is no longer statistically significant. However, Figure 3-4 and Figure 3-5 also show that the reduction in diversity in this region is not fully attenuated, suggesting that other functional motifs may lie in this region that have yet to be identified.

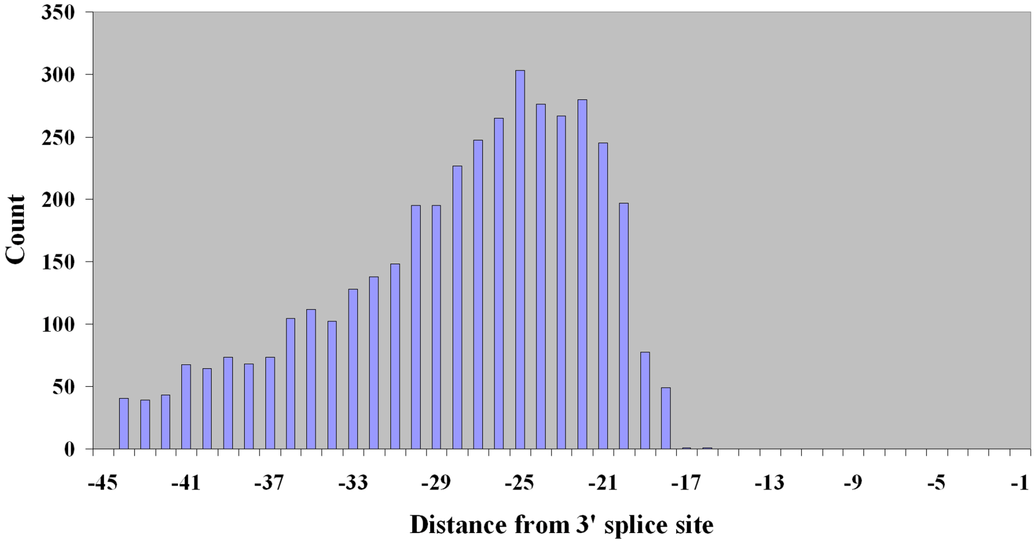

Figure 3-3 Distribution of the position of the branch site adenosine residue from the 3' splice site using the joint PSSM-polyY and HMM prediction method. Distances are relative to the 3' splice site.

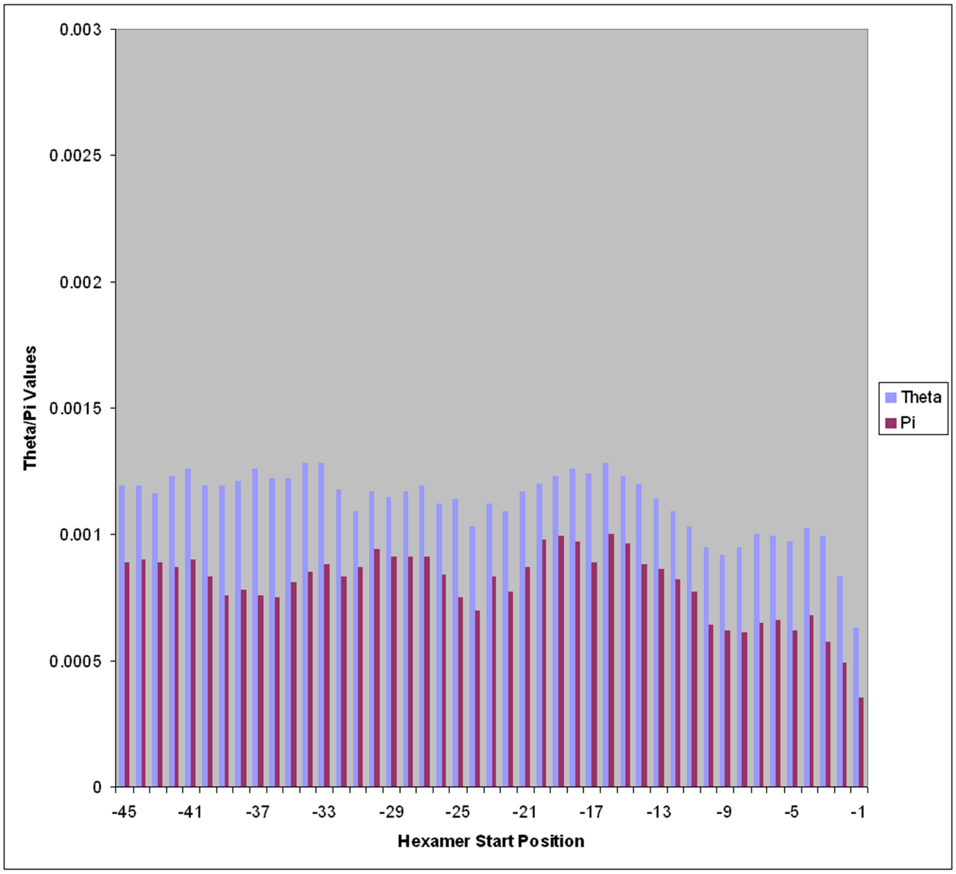

Figure 3-4 Distribution of human polymorphism for hexamers in the 3’ splice site after removing predicted branch sites and using the combined datasets. θ and π values are shown for each hexamer starting at the shown nucleotide positions.

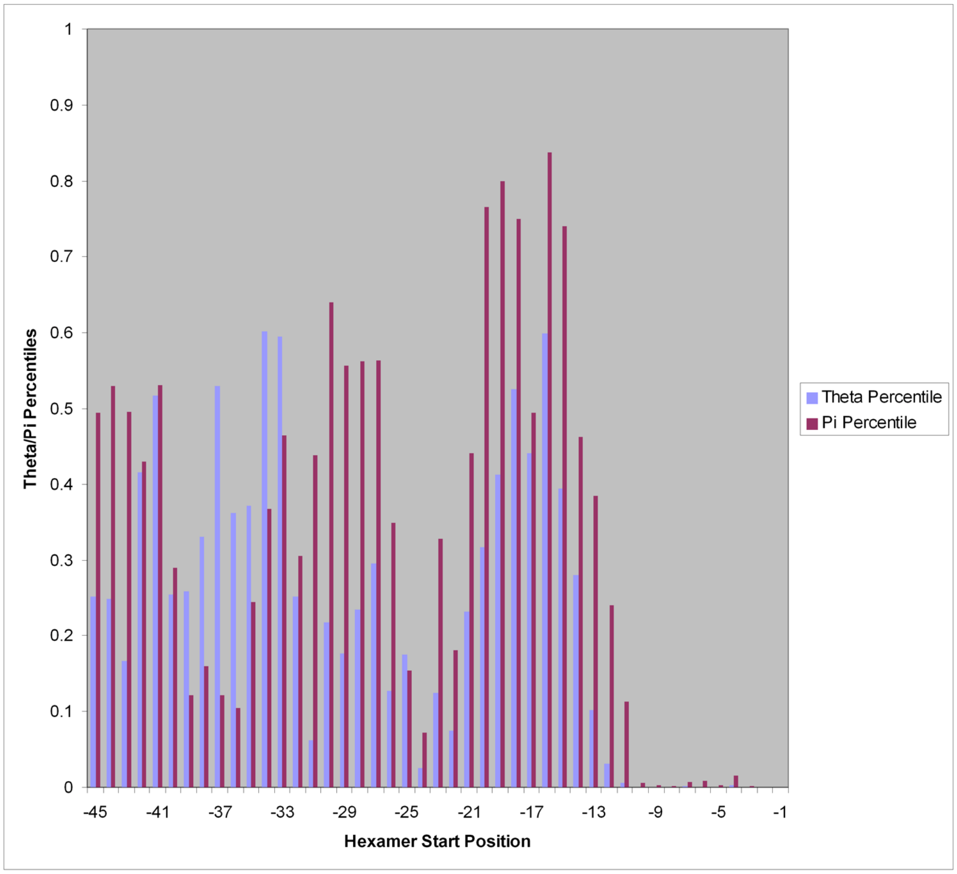

Figure 3-5 Distribution of human polymorphism for hexamers in the 3' splice site after removing predicted branch sites and using the combined datasets. Percentiles are shown for each hexamer starting at the shown nucleotide positions.

3.3.3 Intronic Splicing Enhancer

While the distribution of the branch site adenosine residue (Figure 3-3) clearly explained the reduced polymorphism we previously observed in the 3' region from nucleotide -23 to -27, it did not explain a second region of reduced polymorphism that we observed with a peak from nucleotide -35 to -39 from the 3' splice site (Figure 2-20 and Figure 2-21). To determine if the second peak was due to a distinct functional motif, we applied the genetic variation motif finder to this region. We included nucleotides -31 through -48 for this search, to cover the entire region of reduced polymorphism that we had previously observed. As a positive control, we also searched the region from positions -19 through -35, where we expected to detect the branch site motif.

Results for the region from -19 through -35 showed positive hits to sequences that match a consensus motif corresponding to the known branch-site motif of YURAY (Figure 3-6 and Figure 3-7). On the other hand, the region from -31 through -48 did not reveal any sequences similar to the known branch site motif. However, a positive hit in this region to the unique and distinct motif CCUGG did appear, where the 2nd C had significantly reduced polymorphism levels. This sequence is a subsequence of a known intronic splicing enhancer GGG**CCUGG**G previously identified upstream of the 3' splice site[79].

```
-CUCAC-
    *
--UGACC
    *
-CUGAC-
 *  *
CCUGA--
    *
```

Figure 3-6 Multiple sequence alignment of the predicted functional motifs upstream of the 3' splice site. The multiple sequence alignment of the predicted functional motifs from positions -19 through -35. Positions with reduced genetic variation are represented with an * below.

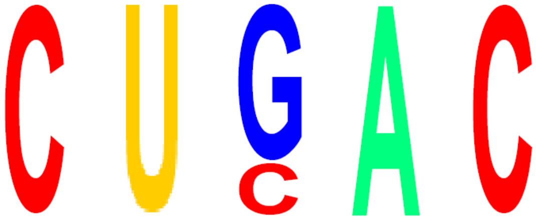

Figure 3-7 Consensus sequence of the predicted functional motifs upstream of the 3' splice site. Consensus sequence generated from the multiple sequence alignment of the predicted functional motifs from positions -19 through -35. The consensus sequence is an excellent match to the branch site consensus sequence.

We hypothesized that the reduced genetic variation we previously observed in this region may be due to the presence of splicing enhancers; mutations in these enhancers have a deleterious effect on proper pre-mRNA splicing. To further explore this hypothesis, we examined the frequency distribution of the CCUGG motif upstream from all 3' splice sites. We found the distribution to be positively skewed and highest at positions -35 to -47 (Figure 3-8), which are within the region of reduced polymorphism previously noted. This distribution supports our conjecture that the reduced genetic variation of nucleotides -31 through -48 are due to the presence of intronic splicing enhancers.

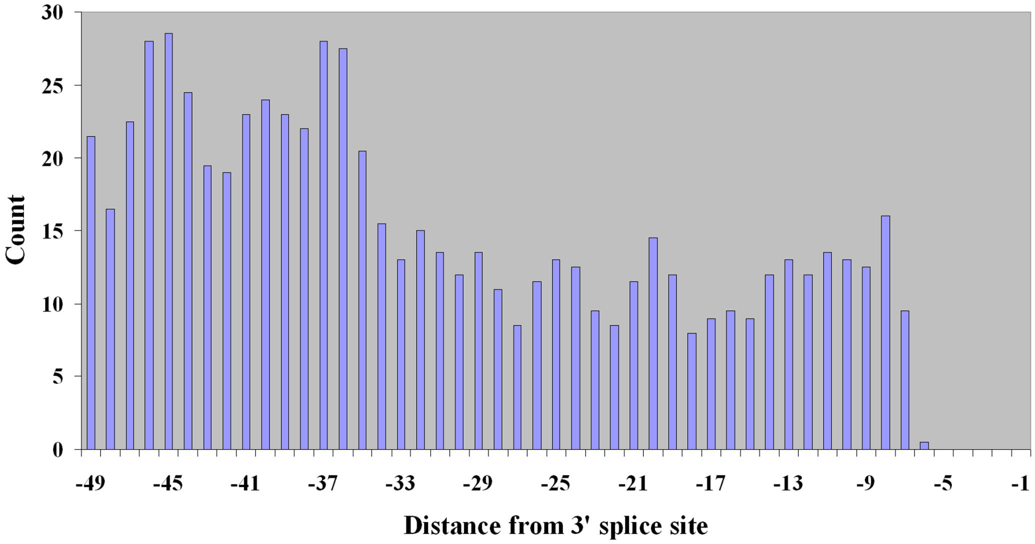

Figure 3-8 Distribution of the position of the sub-sequence, CCUGG, from a known intronic splicing enhancer. Distances are relative to the 3' splice site.

On the 5' side of the intron, we previously observed a peak of reduced genetic variation from nucleotide 21 through nucleotide 30 measured from the 5' splice site (Figure 2-4 and Figure 2-5). We therefore ran the genetic variation motif finder from positions +17 through +34 of the 5' splice site in order to capture all possible pentamers that overlap the originally predicted region from +21 through +30. Results indicated a consensus sequence of GGGCUGGG being the functional motif within this region (Figure 3-9 and Figure 3-10). We found this motif matches a known intronic splicing enhancer $G_3X_{0-4}G_3$[80] which suggests the region from nucleotides +21 through +30 is a generalized location for intronic splicing enhancers. Furthermore, we characterized the frequency distribution of the subsequence, GGCUGG, of this intronic splicing enhancer downstream of the 5' splice site. We find the distribution of this motif is elevated, although not predominant, at positions +21 through +30 (Figure 3-11), with additional peaks at position +11 and +16.

```
---CAGGG
   *
---CUGGG
    *
-GGCUG--
    *
GGGCU---
    *
```

Figure 3-9 Multiple sequence alignment of the predicted functional motifs downstream of the 5' splice site. The multiple sequence alignment of the predicted functional motifs from positions 17 through 34. Positions with reduced genetic variation are represented with an * below.

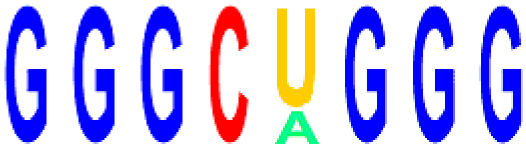

Figure 3-10 Consensus sequence of the predicted functional motifs downstream of the 5' splice site. Consensus sequence generated from the multiple sequence alignment of the predicted functional motifs from positions 17 through 34. The consensus sequence is an excellent match to a previously characterized intronic splicing enhancer.

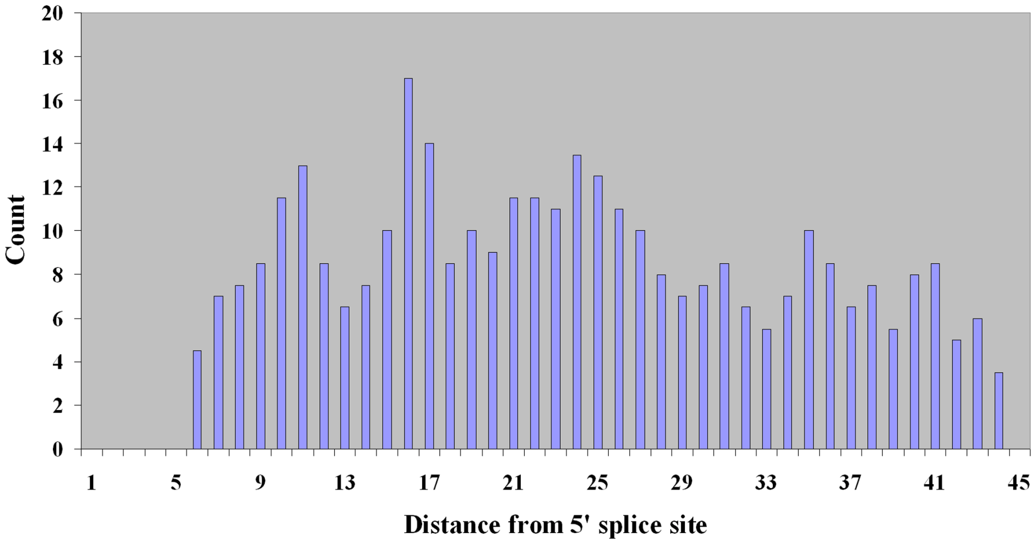

Figure 3-11 Distribution of the position of the sub-sequence, GGCUGG, from a known intronic splicing enhancer. Distances are relative to the 5' splice site.

3.4 Discussion

The combination of genetic variation with transition bias as applied to the PPT not only shows low polymorphism levels that are associated with a functional region, but also insight into the basis for the reduced variation observed in this region. A more in depth analysis showed that purines in the PPT have an increased rate of polymorphism when changed to pyrimidines, suggestive of positive selection. On the other hand, the pyrimidines within the PPT are under negative selection since there is a reduction in both transitions and transversions relative to the control sequences. While the decrease in pyrimidine transversions is not surprising due to the reduced purine content in the PPT, the large decrease in transitions was unexpected and suggests that the specific pyrimidine sequence is also important. This clear preference to increase and maintain the PPT's pyrimidine content suggests that the pyrimidines are essential for its function – an observation that is consistent with previous studies where the binding affinity of

U2AF to the PPT is strengthened with increased pyrimidine content[34,81]. While the elevated polymorphism rate observed in the purines might be considered evidence against functional significance, observing a high polymorphism rate that is specific to only a single class of nucleotides within a specified region, such as the PPT, can reveal the opposite. In this case, the increased genetic variation within the purines may be evidence of a modification of the sequence to one that is better suited for binding to U2AF.

The reduced θ and π values that we observed within the PPT are in agreement with previous suggestions that the PPT is under both positive and negative selection. The reduction of the overall variation found in the PPT reinforces the overall functional significance of the PPT. However, when the nucleotides are divided into classes, the purine transversion rates are higher than those observed for our controls and suggest that within the context of the PPT, these positions are under positive selection. The reduction in variation for pyrimidine sites, for both transitions and transversions, is a strong indicator that these sites are under purifying selection due to their functional importance.

The 3' splice site control that we included in this analysis (positions -27 to -34) also overlaps, to some extent, our branch site localization distribution (although not its peak) and was previously observed to show reduced genetic variation (Figure 2-30 and Figure 2-31). This likely explains some of the reduced variation levels across all the nucleotide classes of this control. On the other hand, the 5' splice site control sequences showed polymorphism levels that would be expected for non-functional controls, although the purines in this region showed modest evidence of reduced genetic variation. A possible explanation for this observation is the higher purine concentration of intronic splicing enhancers that may be located within this region.

The branch site contains two positions with reduced genetic variation – the adenosine occupying the fourth position of the consensus motif, and the position 2 bp upstream occupied by a uracil. Mutations within the uracil residue have been found to cause Ehlers-Danlos syndrome[82], extrapyramidal movement disorder[83], and fish-eye disease[24] which can lead to premature atherosclerosis. These findings, in addition to the known functional properties of the adenosine residue, suggest that the two positions within the branch site are essential for maintaining splicing function. These observations provide further evidence that a nucleotide site's observed genetic variation is a good indicator of its functional significance, even within introns.

Using the branch site prediction methods that we developed and validated for this study, we also examined the location distribution for the branch site adenosine. While the peak location was at position -25, there was a broad spread and there was considerable negative skewness. Thus, while the peak occurred at -25, the inter-quartile range was from position -23 to -32, reinforcing that the region from positions -23 through -32 is a preferred area for branch sites. This region overlaps nearly perfectly with the proximal region of reduced variation we originally observed on the 3' side, past the PPT (Figure 2-20 and Figure 2-21), providing evidence that the branch site is the explanation for the observed reduced genetic variation in this region.

In order to investigate the other regions of reduced genetic variation we observed on both the 3' and 5' sides, we devised a novel motif-finding algorithm based on identifying sequence motifs of reduced genetic variation. We first validated this algorithm by showing that it correctly identified the branch site consensus motif in the region from nucleotides -19 to -35 from the 3' splice site. When we used the same algorithm to identify motifs of reduced genetic variation between nucleotides -31 to -48 the motif CCUGG was identified. This sequence matches a sub-sequence of a previously identified intronic splicing enhancer GGG**CCUGG**G that is upstream of the 3' splice site[79]. Thus, we speculate that the presence of this splicing enhancer within this sequence is the cause of the reduced genetic variation we previously observed (Figure 2-20 and Figure 2-21).

To determine the basis for the reduced genetic variation downstream of the 5' splice site that we observed previously, we implemented the same motif finding algorithm. The consensus sequence that was found, GGGCUGGG, also matches a previously characterized intronic splicing enhancer $G_3X_{0-4}G_3$[80], which is consistent with our prediction that positions +21 through +30 represent a localized region for intronic splicing enhancers that is prevalent amongst most introns from several gene classes. Additionally, this indicates that G_3CUG_3 is the predominant intronic splicing enhancer sequence at this position. The CU positions between the surrounding G_3 motifs were the sites that showed the greatest reduction in genetic variation, suggesting that these positions are important for proper function. However, previous experimental studies have indicated that splicing efficiency is not altered by mutations at these sites whereas mutations in the surrounding G_3 motifs adversely affect function[79,80]. Further analyses will be required to reconcile the difference between the empirical data we observed and the prior experimental data.

In summary we have devised several bioinformatics approaches to further explore the underlying basis for reduced genetic variation observed in specific intronic regions, aside from the well documented splice site sequences. Specifically, we examined the region harboring the PPT on the proximal 3' side, along with two additional regions of reduced variation on the 3' side, with peaks from nucleotides -23 to -27 and -35 to -39. Similarly, we examined a previously identified region of reduced variation at nucleotides +21 to +30 on the 5' side. Our branch site detection method clearly established the region from -23 to -27 on the 3' side as the most likely location for the branch site, explaining the reduced variation observed there. Our motif finding algorithm identified intronic splicing enhancer sequences in the remaining two regions (-35 to -39 on the 3' side, +19 to +32 on the 5' side), providing a likely explanation for the reduced genetic variation observed there.

Our results have important implications for future genetic studies seeking to identify genetic variation associated with human phenotypic traits, such as disease outcomes. Specifically, we have provided a method for characterizing the likely functional significance of genetic variation found in introns. Aside from the splice site sequences (-1 to -6 on the 3' side and +1 to +6 on the 5' side), genetic variation in the branch site nucleotides A and U are obvious targets of potential functional significance. The algorithms we employed show high utility for identifying these nucleotides within all the introns of a gene of interest, or for that matter for all genes. Similarly, the nucleotides we identified within putative splicing enhancers within the introns may also provide annotation that is useful in characterizing intronic variation relevant to disease phenotypes. We would therefore propose that the introns of all genes be analyzed according to the methods we describe here as next generation sequence data become available, to annotate the likely functional intronic variation for all genes.

3.5 Materials and Methods

3.5.1 Genotype Data

The sequence data used for the current analyses has been described in detail previously (2.5.1). Briefly, sequence polymorphism data was acquired from three different sources. The UCSF Pharmacogenetics of Membrane Transporters contains SNP data for 45 human membrane transporter genes[67]; the Seattle SNP database is comprised of 282 genes involved in human inflammatory responses[68]; and the NIEHS SNP database has two sets of genes involved in DNA repair and cell cycle pathways, one containing 385 genes and the other 208 genes[69].

The sequence data from these three sources were used to measure the genetic variation of specific regions using two population genetic parameters, θ and π, which give normalized measures of the expected number of differences between pairs of sequences (2.5.2). The statistical significance of the observed θ and π values was determined by generating distributions of θ and π values from control sequences of similar length from non-functional sites. Empirical p-values were calculated as the proportion of θ and π values from the empirical distributions that were more extreme than the observed values. Further details of these methods are given in a previous section (2.5.4).

3.5.2 Genetic Variation in the Polypyrimidine Tract

The sequence of eight nucleotides from positions -14 through -7 relative to the 3’ splice site’s intron/exon boundary was chosen to study characteristics of the PPT, to maximize the likelihood of inclusion of PPT nucleotides. Position -7 was chosen as the proximal boundary to maximize the length of the sequence being characterized while remaining close to the farthest known distance between the 3’ splice site and a terminal PPT nucleotide, which is position -10[40]. Position -14 was chosen as the distal boundary because the minimum length of the PPT is 10 nucleotides[30] and position -4 of the splice site is the most proximal nucleotide known to never be part of the PPT. Using these boundaries, a single θ and π value was calculated for each intron, within each gene from every database. Additional analyses were based on θ and π values from the individual databases and for each ethnic group across databases. The same measures were calculated comparing the purines and pyrimidines found within the studied sequences to determine whether there are differences in behavior between these two nucleotide classes. Both purines and pyrimidines were further classified into groups depending on the class of the nucleotide change (purine or pyrimidine) in order to determine if certain types of mutations were functionally favored over others in the PPT. The assessment of statistical significance for the observed genetic variation described above was modified to normalize for the nucleotide classes being characterized. For example, if the purine polymorphism levels were measured, then the empirical distribution was generated only from four-fold degenerate (synonymous) purines. When polymorphism levels of sites mutating into specific nucleotide classes were calculated, mutations into the opposite nucleotide class were ignored.

For these analyses, three intronic control regions of the same length (8 nucleotides) were characterized for comparison to ensure that any differences were not due to sample size differences. For the first intronic control, a segment 20 to 13 nucleotides upstream of position -14 on the 3' side (from position -34 through -27) was chosen. On the 5' side, two segments of length 8 corresponding to the same physical locations of the two segments on the 3' side were selected: the first from position 7 through 14, and the other from position 27 to 34.

3.5.3 Identifying the Branch Site: A Hidden Markov Model

The most recently described method for identifying the branch site within the 3' region of the intron is one that uses a position specific scoring matrix, restricted to sequences within 6 nucleotides of predicted PPTs (referred to as PSSM-polyY)[40]. From studies of 46 experimentally validated positive controls, the PSSM-polyY correctly identified 68% of them. The lack of sensitivity was due to ambiguity in the precise location of the PPT.

Using the same 46 sequences with experimentally validated branch sites, we developed a Hidden Markov model (HMM) approach. The HMM was trained using the architecture in Figure 3-12, where individual positions from the branch site to the PPT and the 3' splice site were explicitly modeled. Using this trained model and a 50 bp length sequence as input to the Viterbi algorithm[84], the branch site's location was predicted. The score assigned to the predicted branch site is taken from the matrix generated during the Viterbi algorithm where the maximum score at the end of the predicted branch site is subtracted from its beginning score. Because the transition probabilities in the branch site region being equal to 1, and the maximum scores within each cell in the matrix being converted to log space, the score for the branch site becomes equal to the sum of the log of the emission probabilities for each state in the branch site.

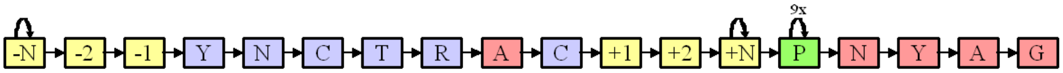

Figure 3-12 Architecture used for the Hidden Markov model. The branch site has 7 states for each of its positions (blue) that surround the conserved adenosine residue (red). Three states were modeled that capture the nucleotide composition surrounding both ends of the branch site (yelllow). Nine positions were explicitly defined for the polypyrimidine tract (green) that were located prior to the four nucleotides of the 3' splice site (red).

To test the sensitivity of the method, a cross-validation training method was used whereby the tested sequence was left out of the training set. This procedure yielded 69.6% correct predictions. Using the PSSM-polyY program[40] to predict branch sites from the 46 control sequences with known branch sites, we obtained 69.6% correct predictions, which matched the performance of the HMM. A ROC curve was generated for each of the two methods to determine how well the scores associated with the branch sites are able to distinguish true positive scores from false positive ones (Figure 3-13 and Figure 3-14). The area under the curve for the PSSM-polyY branch site scores was .866, compared to .672 for the

HMM branch scores. Thus, the PSSM-polyY method alone does a better job discriminating true positive from false positive matches than the HMM alone.

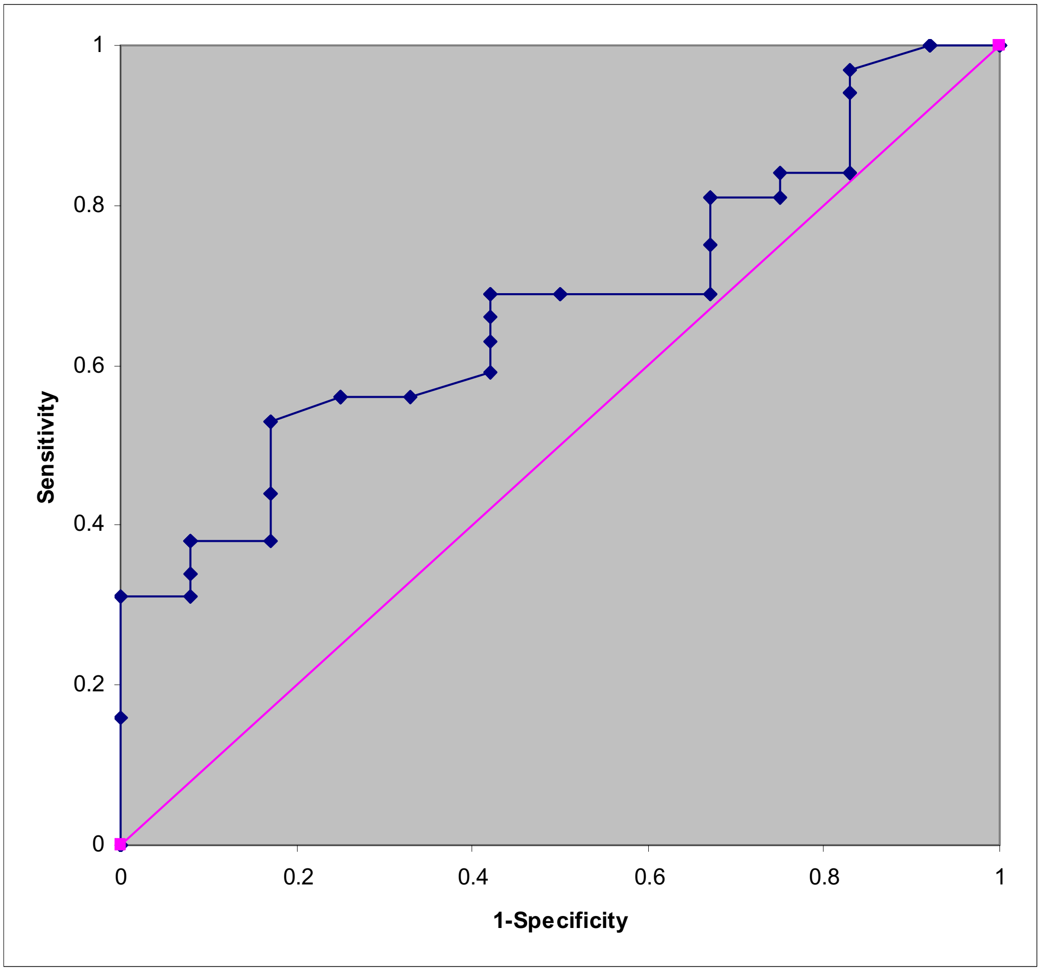

Figure 3-13 ROC curve from using the HMM scores. Plot is generated with the sensitivity and specificity from cross-validation using the positive control sequences at different branch site score cutoffs. The area under the curve is .672 suggesting the branch site score is not an ideal measure to distinguish true positive from false positive scores.

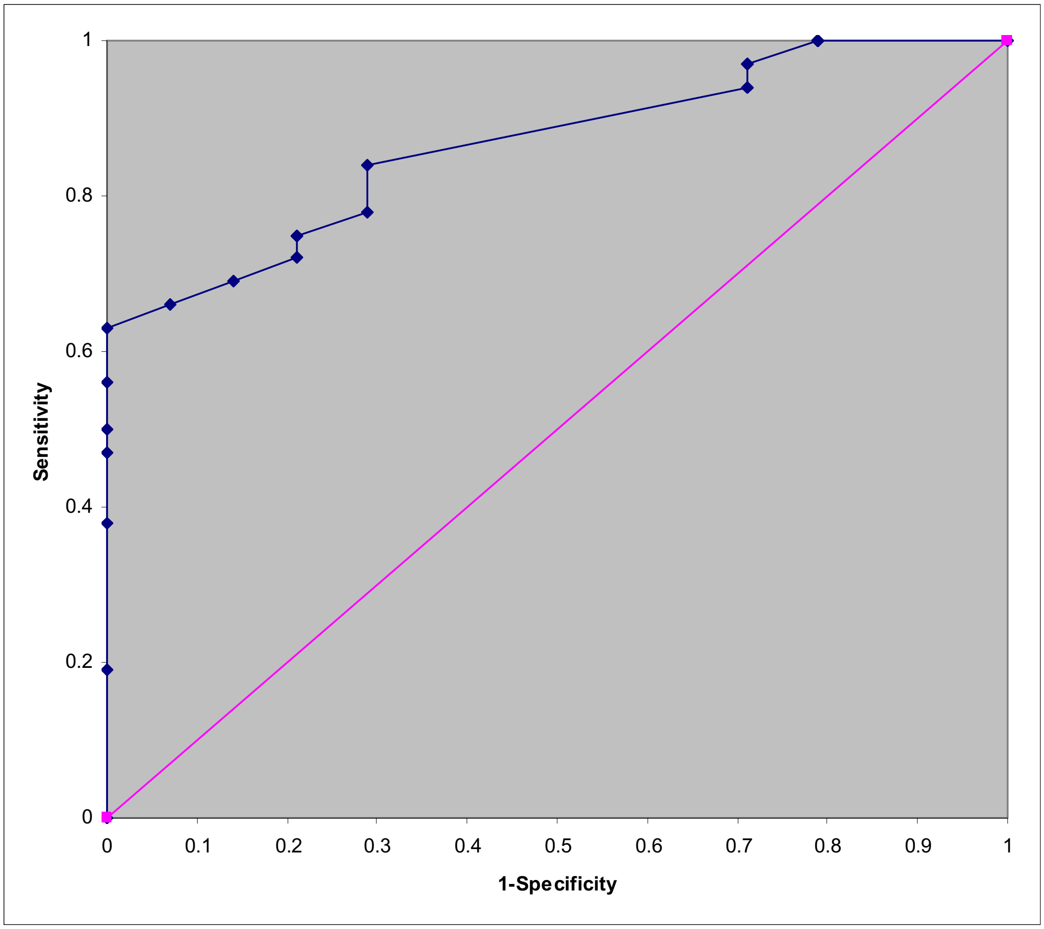

Figure 3-14 ROC curve from using the PSSM-poly scores. Plot is generated with the sensitivity and specificity from cross-validation using the positive control sequences at different branch site score cutoffs. The area under the curve is .866 suggesting the branch site score is well suited to distinguish true positive from false positive scores.

We then considered the two approaches together to determine whether their joint use could improve sensitivity and specificity and ROC characteristics. A direct comparison between the results from the HMM and the PSSM-polyY showed that we could lower the cutoff score for the PSSM-polyY method (to 2.8) if we required the HMM to predict the same branch site, to maintain the same specificity. With this double checking between the two different methods, we were able to increase the sensitivity of the PSSM-polyY method from .63 to .78

without any loss of specificity (Figure 3-15). Therefore, the two methods were combined for final branch site prediction.

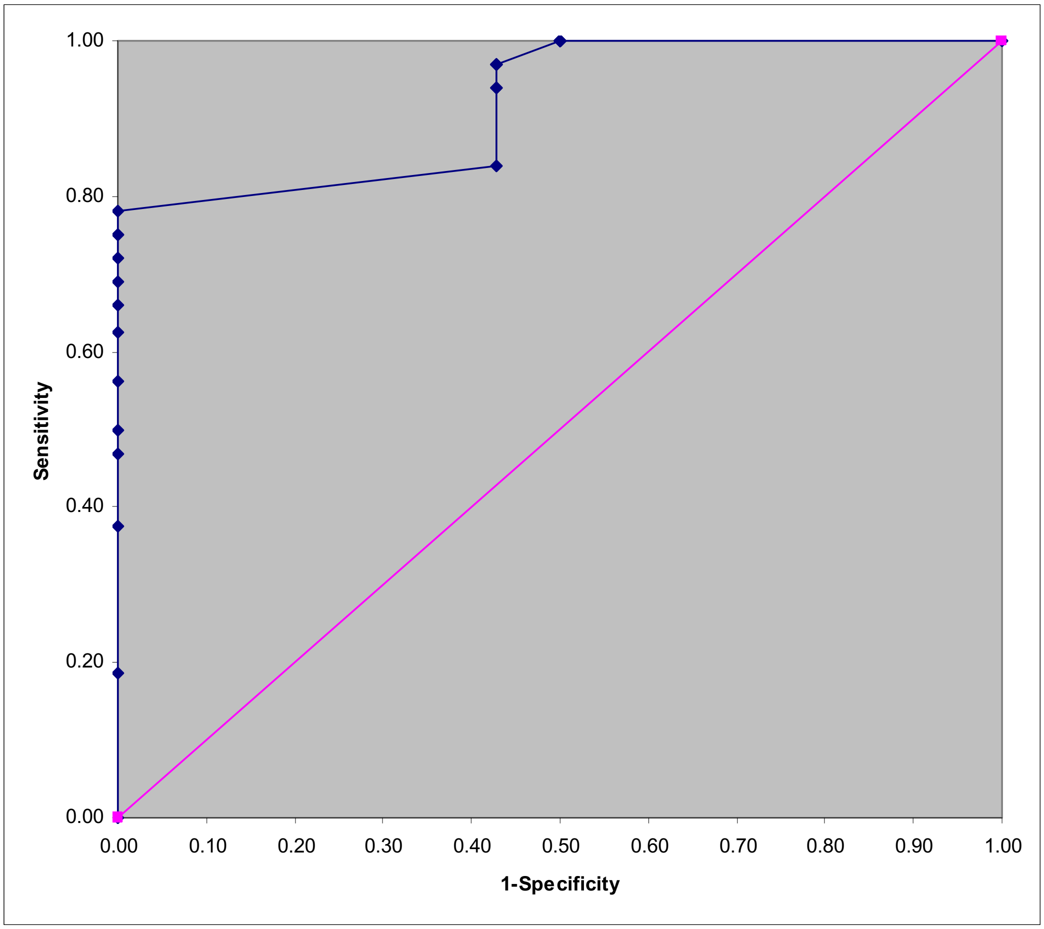

Figure 3-15 ROC curve from using the joint PSSM-polyY and HMM scores. Plot is generated with the sensitivity and specificity from cross-validation using the positive control sequences at different branch site score cutoffs. The area under the curve is .917 suggesting the combined method is very well suited to distinguish true positive from false positive scores.

3.5.4 Genetic Variation Motif Finder

The premise for creating a new functional motif finder was that a motif should show signs of reduced genetic variation within its functional positions, regardless of its relative frequency of occurrence. On this basis, we searched for functional motifs by: (1) Selecting the largest n-mer size possible that can still yield adequate statistical power when measuring genetic variation; for our datasets, pentamers were used since hexamers were not sufficiently abundant to yield adequate power; (2) Selecting a region for analysis, such as positions -19 through -36 upstream of the 3' splice site where we previously observed reduced genetic variation; (3) Within the chosen region, grouping all possible overlapping pentamers from all introns within every gene into groups based on the unique pentamer sequence; (4) Using the available SNP information from the three databases, measuring θ and π for each individual position for every unique pentamer group and keeping only those which have at least one position where both θ and π are below a 5% cutoff; (5) Hierarchically clustering pentamer sequences based on a sequence dissimilarity metric with complete linkage (distance between two clusters is the distance between the farthest elements from each cluster) criteria; and (6) Selecting the hierarchy which separates clusters by the largest distance under two dissimilarity units (number of mismatches between the aligned consensus sequences from each cluster) and performing a multiple sequence alignment within each cluster.

4 Measuring Genetic Variation in Coding Regions

While the primary goal of this thesis is to characterize the genetic variation in non-coding regions, its success within non-coding regions prompted the question if we could improve the characterization of functional elements in coding regions. Therefore, this chapter will present our analyses of the coding regions neighboring the splice sites since their positions are static and they are known to interact with snRNPs during splicing. As such, there is an expectation that the sites from these regions would show highly reduced genetic variation. However, characterizing the functional components within coding regions will be more difficult since there is already selective pressure to conserve a gene's amino acid sequence, effectively overlapping any reduced genetic variation due to snRNP binding.

4.1 Introduction

The identification of functional elements within coding and noncoding sequence is of importance for understanding how mutations can lead to human disease. One such method that has proven useful for identifying functional regions in noncoding sequence is the identification of sites with reduced genetic variation (2.4 and 3.4). The foundation for this method is based on the idea that positions with decreased polymorphism levels are suggested to be under purifying selection and are therefore likely functional candidates. Using this method, we have shown that several known noncoding functional splicing elements including the 5’ splice site, 3’ splice site, polypyrimidine tract, branch site, and putative intronic splicing enhancer locations all show reduced polymorphism levels (2.3.2, 2.3.3, 3.3.1, 3.3.2, and 3.3.3). Given the success of these previous analyses, we expand on using human genetic variation within coding sites to identify regions of functional importance. While most coding sites are nonsynonymous (nondegenerate) and are therefore considered functional, synonymous (four-fold degenerate) sites in coding regions are commonly accepted as nonfunctional since mutations at these sites will not change the amino acid sequence[56]. Therefore, the recognition of functional synonymous variants is a difficult task, and for this reason, we aim to use human genetic variation within coding sites to identify functionally important synonymous sites.

Previous analyses in coding regions have shown that single nucleotide polymorphisms (SNPs) located in nonsynonymous positions occur less frequently, in particular if mutations at those sites lead to non-conservative amino acid changes[17] (Table 2-1). Similar studies reinforced this observation whereby sequence variants with nonsynonymous substitutions have lower allele frequencies[12]. These observations are consistent with selective pressure acting to maintain the amino acid sequence encoded by a gene, particularly at positions where mutations cause dramatic changes in the properties of the amino acid. However, there are additional properties within a gene other than its amino acid sequence that are important for its function such as mRNA secondary structure stabilization[56,57], regulatory splicing motifs like exonic splicing enhancers and silencers[58], and codon usage bias[59]. Here we present a study on using genetic variation within coding regions that allows for a more in depth characterization of alternative functional constraints in a gene that are not related to amino acid conservation. The regulation of RNA splicing is a known functional constraint within coding sites that is unrelated to sequence conservation. Several regions within introns and exons are necessary for RNA splicing and act as recognition and binding sites for the spliceosome – the complex responsible for intron removal and exon ligation. The spliceosome is composed of individual subunits known as small nuclear ribonucleoprotein particles (snRNPs), which primarily bind to motifs found within introns. However, some snRNPs, like U1 and U5, also bind to sequences within exons. U1 binds to introns in the 5' splice site at positions 1, 2, 3, 5, and 6 while simultaneously binding to the last two positions of the neighboring 3' exon end[34]. U5 is an integral part of splicing that joins the 5' and 3' splice sites with one another by binding to the last three positions of the 3' exon end and the first two positions of the 5' exon end[34,67]. Within this study, we use human polymorphism data to identify exon variants that do not change the amino acid sequence yet may influence gene function through the regulation of RNA splicing.

4.2 Results

4.2.1 All Sites

We examined the genetic variation at both 5' (Figure 4-1 and Figure 4-2) and 3' (Figure 4-3 and Figure 4-4) exon ends at each individual nucleotide position based on distance from the intron/exon boundary. Every position for both the 5' and 3' exon ends consistently has low genetic variation except for position -1 in the 3’ exon end (Figure 4-3 and Figure 4-4). The reduced genetic variation seen within these positions is expected due to the functional requirement of maintaining a gene’s amino acid sequence.

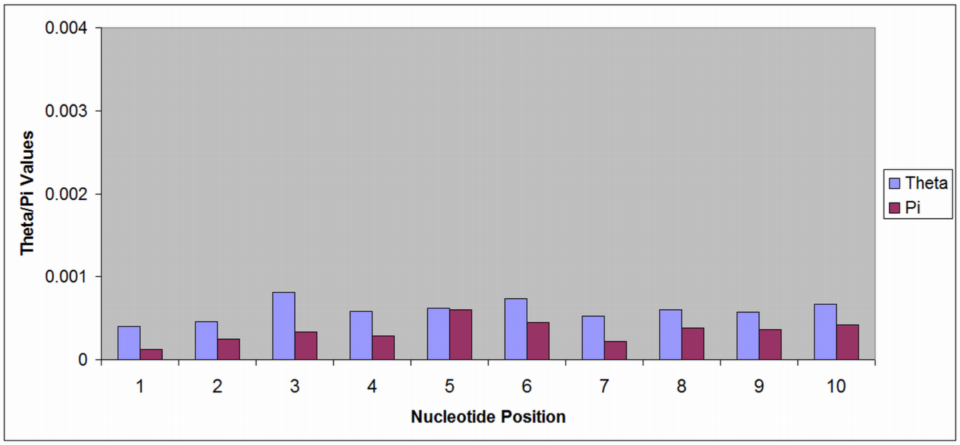

Figure 4-1 Distribution of human polymorphism within the 5’ ends of all exons using the combined datasets. θ and π values are shown for each position.

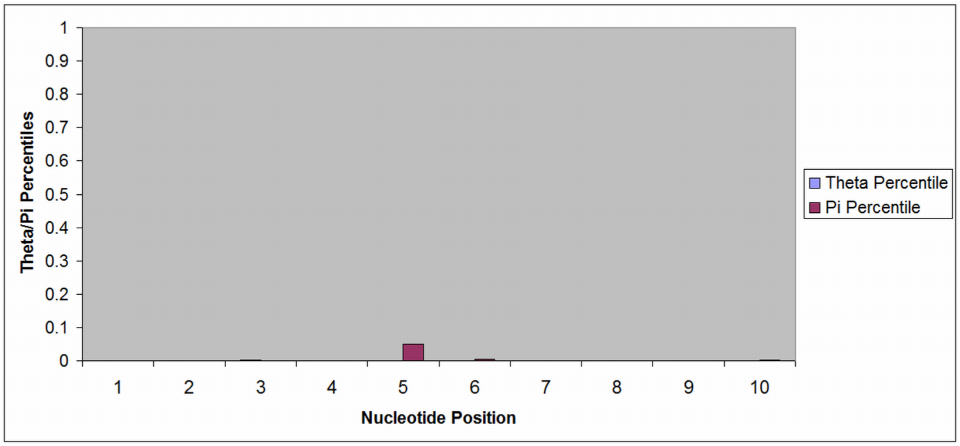

Figure 4-2 Distribution of human polymorphism within the 5’ ends of all exons using the combined datasets. Percentiles are shown for each position.

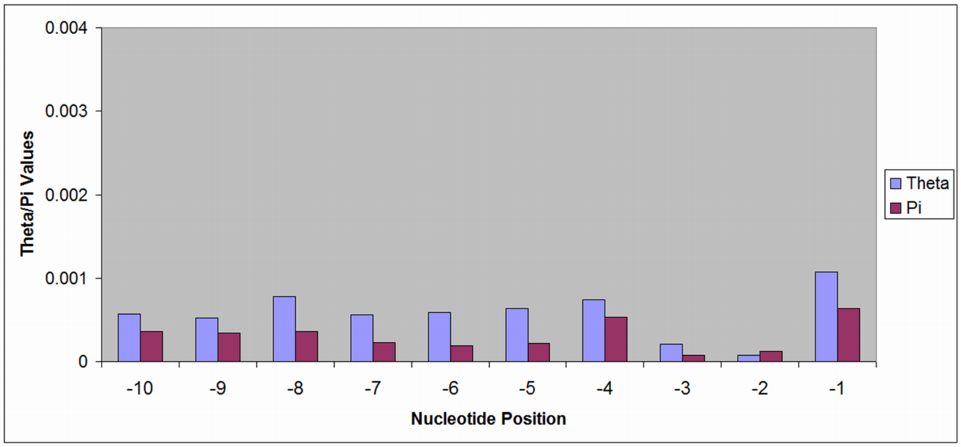

Figure 4-3 Distribution of human polymorphism within the 3’ ends of all exons using the combined datasets. θ and π values are shown for each position.

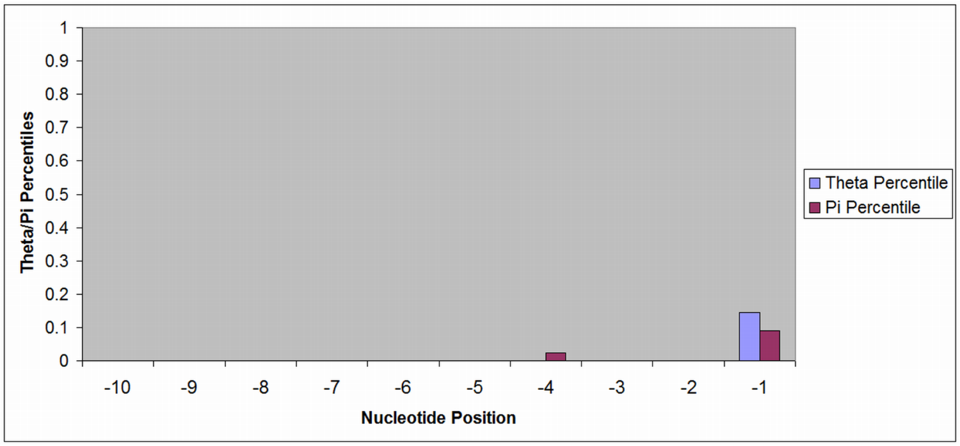

Figure 4-4 Distribution of human polymorphism within the 3’ ends of all exons using the combined datasets. Percentiles are shown for each position.

4.2.2 Nondegenerate Sites

To further investigate the properties of the 5’ (Figure 4-5 and Figure 4-6) and 3’ (Figure 4-7 and Figure 4-8) exons, we characterized their genetic variation using only nondegenerate nucleotides. Nondegenerate positions are those in which any nucleotide substitution will cause a mutation in a gene’s amino acid sequence. Nondegenerate polymorphism levels are highly reduced at all positions, similar to our first results using all nucleotide sites. While position -1 in the 3’ exon (Figure 4-7 and Figure 4-8) is low (2.5% range for both θ and π), it is higher than all other nondegenerate sites in both 5’ and 3’ exons.

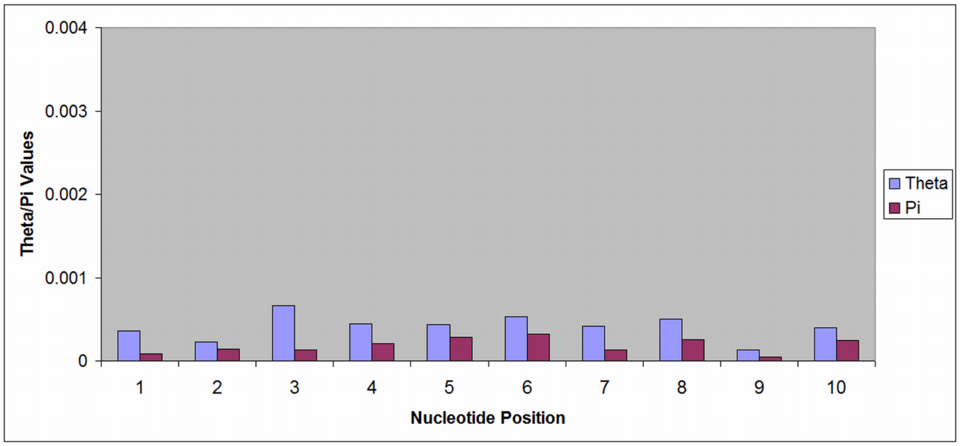

Figure 4-5 Distribution of human polymorphism within the 5’ ends of all exons using the combined datasets and only nondegenerate sites. θ and π values are shown for each position.

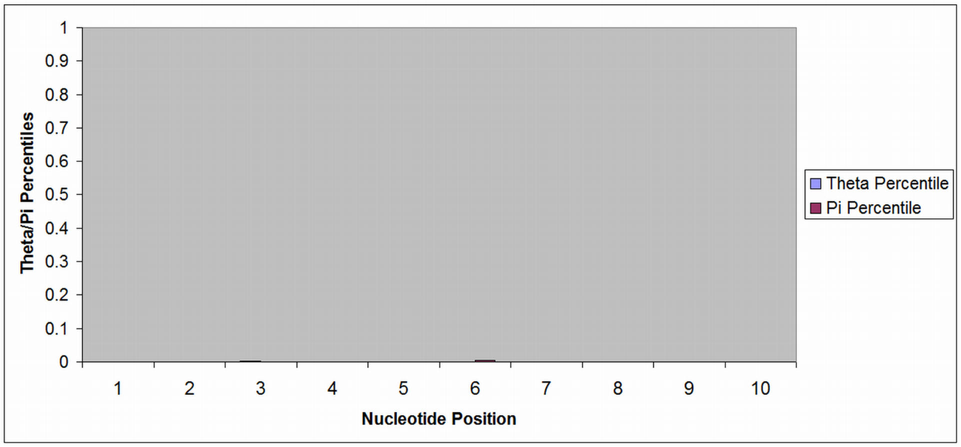

Figure 4-6 Distribution of human polymorphism within the 5’ ends of all exons using the combined datasets and only nondegenerate sites. Percentiles are shown for each position.

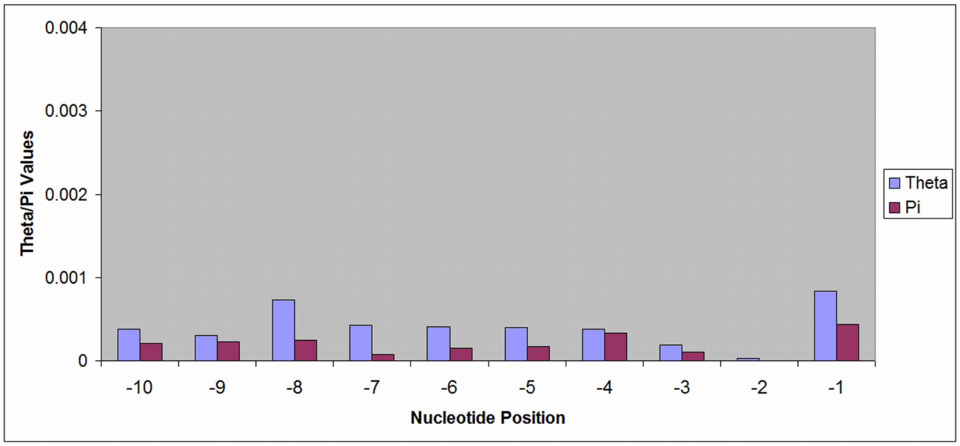

Figure 4-7 Distribution of human polymorphism within the 3' ends of all exons using the combined datasets and only nondegenerate sites. θ and π values are shown for each position.

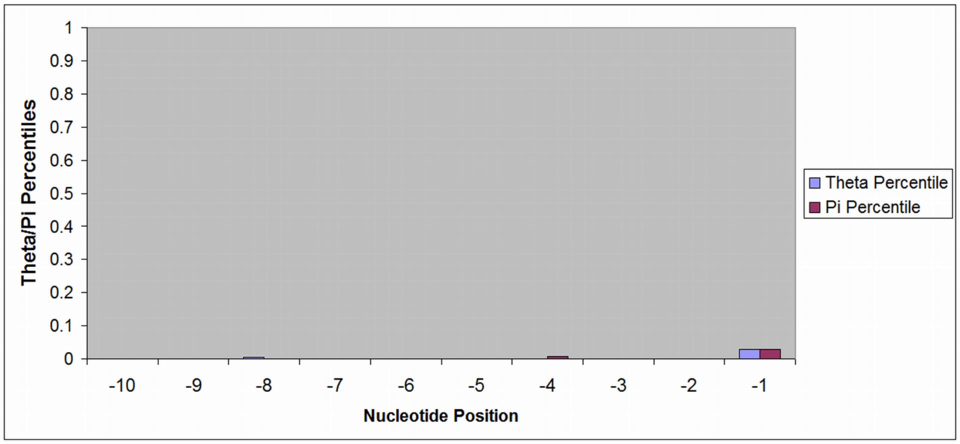

Figure 4-8 Distribution of human polymorphism within the 3' ends of all exons using the combined datasets and only nondegenerate sites. Percentiles are shown for each position.

4.2.3 Four-fold Degenerate Sites

Similar to the analysis described above using only nondegenerate sites, we repeat the same experiment using four-fold degenerate (synonymous) sites. These are sites where all three mutations will not change a gene's amino acid sequence. Characterization of the genetic variation of the 3' exon end (Figure 4-9 and Figure 4-10) shows that positions -2 and -3 have significantly reduced polymorphism levels for both θ and π. Additionally, positions -5 (low π) and -7 (low θ) have low variation when measuring only one of the population genetic parameters. Position -1 continues to show higher genetic variation than other coding sites as has been shown previously in Figure 4-3 and Figure 4-4.

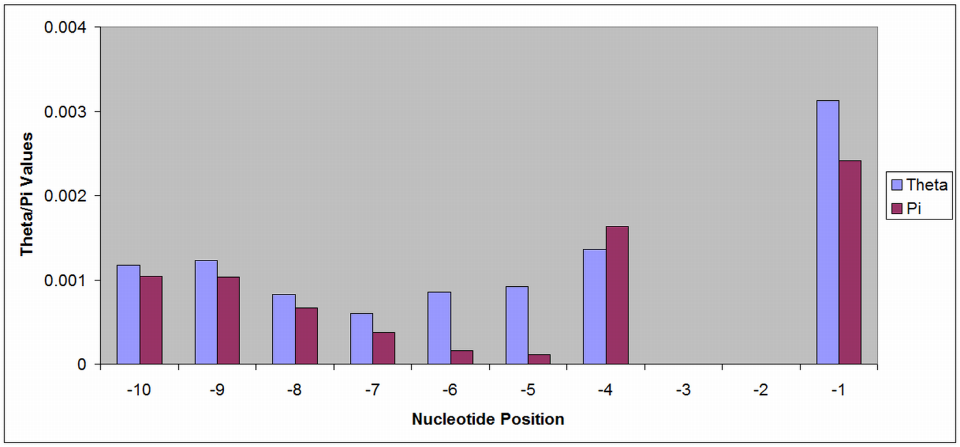

Figure 4-9 Distribution of human polymorphism within the 3' ends of all exons using the combined datasets and only four-fold degenerate sites. θ and π values are shown for each position.

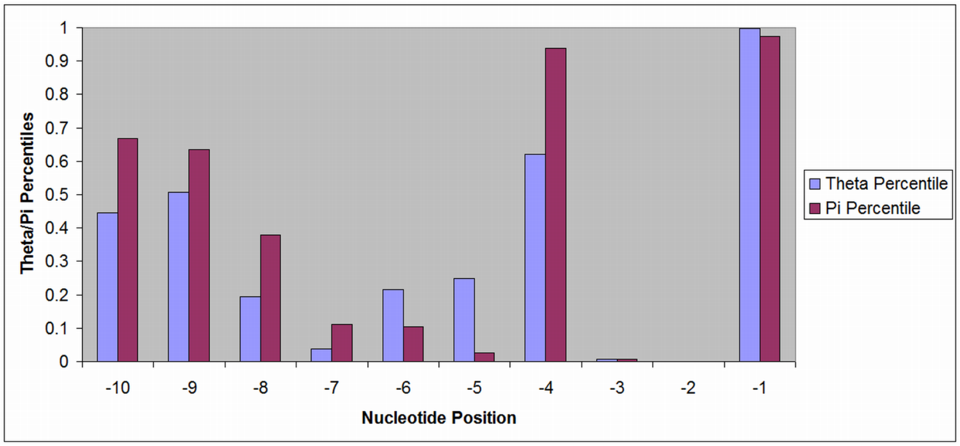

Figure 4-10 Distribution of human polymorphism within the 3' ends of all exons using the combined datasets and only four-fold degenerate sites. Percentiles are shown for each position.

4.2.4 3' Exon Position -1

In a previous section, we found we were able to identify the functional properties and behavior of the polypyrimidine tract, an important intronic motif required for intron splicing (3.3.1). We measured the genetic variation of the polypyrimidine tract when grouped by nucleotide class (pyrimidines - Y and purines - R) and by specific nucleotide class mutations (Y→R, Y→Y, R→Y, R→R). The study showed not only which nucleotide groups are both resistant and prone to mutations, but also indicated that some regions select for mutations into specific nucleotide classes. The inferred function and behavior of the polypyrimidine tract using those results matched well with known experimental results from the literature. Given the success of this previous method, we applied them to position -1 of the 3’ exon end (Table 4-1) to characterize its functional properties. The overall nondegenerate sites show low polymorphism levels (Table 4-1) as was noted previously (Figure 4-7 and Figure 4-8); however, nondegenerate pyrimidines have much higher polymorphism levels than purines. Comparatively, the four-fold degenerate sites at position -1 show higher levels of genetic variation regardless of nucleotide class. The difference between the polymorphism levels in purines and pyrimidines within nondegenerate sites shows that those pyrimidines within nondegenerate sites are missing the normal selective pressure that reduces genetic variation levels.

Table 4-1 Summary of π and θ values and their corresponding percentiles for the different nucleotide classes within position -1 from all 3' coding exons.

					Pi				
Sites	**Overall**	**R**[a]	**R_{TI}**[b]	**R_{TV}**[c]	**$R_{TI:TV}$**[d]	**Y**[e]	**Y_{TI}**[f]	**Y_{TV}**[g]	**$Y_{TI:TV}$**[h]
Nondegenerate	4.40	1.90	1.80	0.10	18.00	31.60	30.30	1.30	23.31
Four-fold Degenerate	24.10	20.60	20.60	0.10	206.00	42.10	6.90	35.20	0.20

					Θ				
Sites	**Overall**	**R**[a]	**R_{TI}**[b]	**R_{TV}**[c]	**$R_{TI:TV}$**[d]	**Y**[e]	**Y_{TI}**[f]	**Y_{TV}**[g]	**$Y_{TI:TV}$**[h]
Nondegenerate	8.40	4.80	3.90	0.90	4.33	48.20	38.80	9.40	4.13
Four-fold Degenerate	31.30	28.20	28.20	0.10	282.00	46.80	31.20	15.60	2.00

Pi Percentiles

Sites	Overall	R[a]	R_{TI}[b]	R_{TV}[c]	Y[e]	Y_{TI}[f]	Y_{TV}[g]
Nondegenerate	0.029	0.000	0.001	0.005	0.989	0.997	0.699
Four-fold Degenerate	0.975	0.862	0.943	0.268	0.981	0.812	0.995

Θ Percentiles

Region	Overall	R[a]	R_{TI}[b]	R_{TV}[c]	Y[e]	Y_{TI}[f]	Y_{TV}[g]
Nondegenerate	0.028	0.000	0.000	0.005	1.000	1.000	0.949
Four-fold Degenerate	0.998	0.956	0.993	0.264	0.989	0.963	0.965

[a]R: Purines, [b]R_{TI}: Purine Transitions, [c]R_{TV}: Purine Transversions, [d]$R_{TI:TV}$: Purine transition/transversion ratio, [e]Y: Pyrimidines, [f]Y_{TI}: Pyrimidine Transitions, [g]Y_{TV}: Pyrimidine Transversions, [h]$Y_{TI:TV}$: Pyrimidine transition/transversion ratio

4.3 Discussion

Using standard measures of nucleotide variation within coding regions showed significantly reduced genetic variation levels that reflect the importance of amino acid conservation within genes (Figure 4-1 through Figure 4-4). By dividing these regions into synonymous and nonsynonymous sites however, we are able to gain more insight into the characteristics of the observed genetic variation levels within coding regions. Significantly reduced genetic variation within nonsynonymous sites are expected due to their functional constraints to maintain the amino acid sequence encoded by a gene. Since amino acid sequence-based conservation constraints do not exist within synonymous sites, reduced polymorphism levels at these sites indicate the presence of alternative functional restrictions such splicing motifs, mRNA secondary structure stabilization, and codon biases. For example, positions -2 and -3 of the 3' splice site show significantly reduced polymorphism levels at both synonymous (Figure 4-9 and Figure 4-10) and nonsynonymous (Figure 4-7 and Figure 4-8) sites. This suggests that these two positions are important for both amino acid conservation and another functional constraint. As we described previously in this study, position -2 is important for binding to the U1 and U5 snRNPs during pre-mRNA splicing and position -3 is important for binding to the U5 snRNP, reinforcing the additional functional constraint implied by the reduced synonymous polymorphism levels.

To further characterize the increased levels of genetic variation of position -1 in the 3' exon (Figure 4-3 and Figure 4-4), we divided this site into synonymous and nonsynonymous groups to look for differences in the properties of nucleotide classes. The genetic variation of the nonsynonymous group is significantly low (Table 4-1), agreeing with the expected functional constraints of amino acid conservation. Furthermore, the genetic variation of the synonymous group is high, also agreeing with the decreased functional importance expected at synonymous sites[56]. Since the observed levels of genetic variation in both these groups did not explain our original observation, we further divided the nonsynonymous group into purines and pyrimidines. We then found the purines to have significantly low levels of genetic variation, while the pyrimidines showed uncharacteristically high levels since these positions should theoretically still have a selective pressure to maintain the amino acid sequence encoded by a gene. Position -1, which contains predominantly purines (90.5% purines across all datasets), is part of the 5' splice site motif and is known to bind to the U1 and U5 snRNPs during pre-mRNA splicing[33,34], suggesting the basis of the increased genetic variation observed within nonsynonymous pyrimidines may be due to a delicate balance between creating a more typical splice site motif at the expense of generating nonsynonymous mutations. This result however, is only tentative since there were only 5,122 datapoints for this position across all datasets, making the analyses for any positions other than the nonsynonymous purines less reliable (79.3% of datapoints are nonsynonymous purines).

Transition bias analysis is a model that provides insight into the process of nucleotide substitution and is defined as the ratio (TI:TV) of the rate of transitions (Y→Y, R→R) to the rate of transversions (Y→R, R→Y)[76]. Assuming no substitution preferences, the TI:TV ratio should be equal to .5 due to there being twice as many options for a transversion than a transition; however, it has long been recognized that transitions occur more frequently than transversions[76]. Typical values in coding regions range from 1.86 to 3.89 and from 1.99 to 3.22 within intron regions (3.2.3 and 3.3.1). It has been noted before that the reasons for the increased transitions levels is primarily due to the fact that they cause less amino acid substitutions than transversions and if they do, the resulting mutations tend to maintain the chemical properties of the native amino acid[76]. Due to our success using this model as described previously (3.2.3 and 3.3.1), we applied it again to position -1 of the 3' exon (Table 4-1) to determine if the mutational properties of this position exhibit any deviations from expected values. When these ratios are measured using synonymous and nonsynonymous positions (Table 4-1), we observe that the results are higher than the observed values we described previously (3.2.3 and 3.3.1). Given that the general observation for increased transition levels is primarily due to minimizing harmful amino acid changes, the measured TI:TV ratios specifically within nonsynonymous groups in this study are likely inflated since mutations at nonsynonymous sites will always change the translated gene sequence. Synonymous sites on the other hand do not have these constraints, which makes the observed TI:TV ratios within synonymous purine sites that are larger than 200 outside the normal range of observed values. The main cause of these large values appears to be due to decreases in frequency of purine to pyrimidine changes as we previously described. A caveat that should be reiterated is that the TI:TV ratios for groups other than the nonsynonymuos purines should not be considered very reliable due to their smaller sample sizes.

These data suggest that measurements of genetic variation within coding regions can be used to differentiate regions that are functional due to amino acid conservation from those with alternate functions such as splicing motifs. The normal genetic variation measurements within coding regions will be highly reduced due to the functional constraints of maintaining a gene sequence. However, dividing coding regions into synonymous and nonsynonymous sites can be useful for distinguishing different functional constraints within regions. Nonsynonymous sites with reduced polymorphism levels are expected due to the functional constraints of gene sequence conservation. On the other hand, increased nonsynonymous polymorphism levels, such as those seen in the pyrimidines at position -1 of the 3’ exon end, suggest a deviation from the normal sequence conservation constraints. Characterizing the genetic variation of synonymous sites can then be useful since they lack the sequence conservation constraints. This means that reduced polymorphism levels at synonymous sites indicate the presence of alternative functional elements such splicing motifs, mRNA secondary structure stabilization, and codon biases.

4.4 Materials and Methods

DNA sequences and analytical methods used in this study were implemented as described in detail in earlier sections (2.5.1 and 3.5.1). Briefly, SNP data was collected from three different sources. The UCSF Pharmacogenetics of Membrane Transporters houses SNP data for 45 human membrane transporter genes that were generated by sequencing 100 African Americans, 100 Caucasians, 30 Asians, 10 Hispanics, and 7 Pacific Islanders[67]. The Seattle SNP database is composed of 282 genes involved in human inflammatory responses that were generated by sequencing 24 African Americans and 23 Caucasians[68]. The NIEHS SNP database is composed of SNP data for 385 genes involved in DNA repair and cell cycle pathways that were generated by sequencing 90 individuals that were representative of the U.S. population and include European-Americans, African-Americans, Mexican-Americans, Native-Americans, and Asian-Americans (NIEHS phase 1) in an undisclosed proportion. An additional set from the NIEHS SNP database was also used that is composed of SNP data from 208 genes from sequencing 95 individuals: 27 African-Americans, 22 Caucasians, 22 Mexican-Americans, and 24 Asian-Americans (NIEHS phase 2)[69].

The SNP data from these databases is used to measure the genetic variation of specific regions using two population genetic parameters, θ and π, which give measures of nucleotide polymorphism levels that are normalized for both sequence length and population sample size. The statistical significance of the measured variation is estimated by calculating the statistical probability of observing the measured θ and π values within empirical distributions of random θ and π values taken from non-functional sites (2.5.2 and 2.5.4).

The selection of nucleotide regions for analyses was dependent on the location of known exon sequences that are bound to by the spliceosome. As described earlier, the U1 and U5 snRNPs are known to bind to the 5' and 3' exon ends immediately neighboring the intron/exon boundaries (2.2.2). For this reason, we measured the genetic variation for every individual position from the first and last 10 nucleotides of every exon end (Figure 4-11) similar to how intron ends were characterized previously (2.5.5). Numbering for the nucleotide positions of the exon ends is relative to the nearest intron/exon boundary as shown in Figure 4-11.

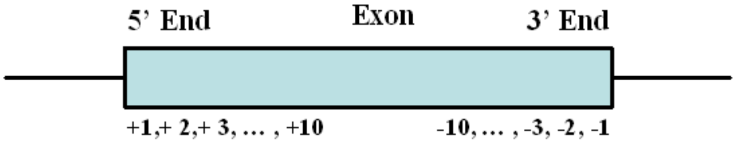

Figure 4-11 Exon regions selected for analysis and numbering system used. Only the first and last 10 nucleotides of each exon end were used in this analysis.

Conclusions

The primary goal of this thesis was to investigate if human polymorphism data can be used to characterize functional elements within non-coding regions. Using publicly available human gene sequence data we found that several known and novel non-coding functional elements display levels of genetic variation that are significantly low compared to our empirical non-functional distributions. Among these, known functional elements with relatively static localization sites, like the splice sites and polypyrimidine tract, provided validation that functional non-coding motifs exhibit reduced polymorphism levels. The analyses done to characterize the properties of the splice sites and polypyrimidine tract revealed additional regions that appeared to contain intronic functional elements, prompting further studies. Using computational methods to characterize these regions, we found that the reduced genetic variation was caused by the presence of the branch site and intronic splicing enhancers.

The measurement of a region's genetic variation can be used as a binary classification method for labeling a region as functional or non-functional. However, we found that through the characterization of sites that are divided into different groups, such as pyrimidines and purines, we can identify more specific properties of a region's putative functional roles. For instance, the polypyrimidine tract showed not only high purine transversion levels but also reduced pyrimidine polymorphism levels, agreeing with the functional importance of pyrimidines for proper binding to the U2AF snRNP. Furthermore, we found that measuring the genetic variation of synonymous and nonsynonymous groups within coding regions was useful for inferring functional properties that are not associated with the conservation of a protein's amino acid sequence. Synonymous sites with decreased levels of genetic variation and nonsynonymous sites with increased levels were found to contradict the expected distribution of genetic variation at those positions (Table 2-1). We found that sites displaying both of these phenotypes were positions that are important for the binding of snRNPs during mRNA splicing, suggesting that our methods are usable in both coding and non-coding regions.

The primary limitation of our analyses was the availability of sequencing data. This constraint was due to the fact that only a limited number of genes across a small number of individuals were sequenced and publicly available. While we were able to collate data within introns to bypass this limitation, our conclusions were the result of the consensus properties across several individuals and genes as opposed to identifying specific functional sites. Once more sequencing information for the human population becomes available, these types of analyses can be refined and implemented without the need to combine data from several similar regions. This would theoretically make it possible to simply measure the polymorphism within a single position across a very large number of individuals to predict if that site is functionally important. Nonetheless, we have shown a proof of principle demonstrating that the genetic variation between individuals can likely be used to characterize regions of functional importance across the entire human genome.

As mentioned previously, the availability of more sequencing data would facilitate the application and use of the methods found in this thesis. Future genetic studies seeking to associate phenotypic traits, including diseases, with genetic variants can be improved by implementing the methods discussed in this thesis. Current studies however, could use the algorithms we have developed to identify potential disease causing nucleotides specifically within introns. Any SNPs that are located within the canonical splice site or polypyrimidine tract regions, predicted branch sites, or hypothetical intronic splicing enhancers may be useful for associating intronic variation with disease phenotypes.

References

[1] http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome

[2] Venter, JC, et al. (2001) The sequence of the human genome. Science, 291, 1304-1351.

[3] Ashurst, JL, and Collins, JE. (2003) Gene Annotation: Prediction and Testing. Annu. Rev. Genomics Hum. Genet, 4, 69-88.

[4] Davuluri, RV, Grosse, I, and Zhang, MQ. (2001) Computation identification of promoters and first exons in the human genome. Nature, 29, 412-417.

[5] Stein, L. (2001) Genome Annotation: From sequence to biology. Nature Reviews Genetics, 2, 493-503.

[6] Zhang, MQ. (2002) Computational prediction of eukaryotic protein-coding genes. Nature Genetics Reviews, 3, 698-709.

[7] Hardison, RC. (2000) Conserved Noncoding Sequences are Reliable Guides to Regulatory Elements. Trends Genet. 16, 369-372.

[8] Ganley, ARD, Hayashi, K, Horiuchi, T, and Kobayashi, T. (2005) Identifying Gene-Independent Noncoding Functional Elements in the Yeast Ribosomal DNA by Phylogenetic Footprinting. PNAS, 102, 11787-11792.

[9] Dermitzakis, ET, Reymond, A, Lyle, R, Scamuffa, N, Ucla, C, Deutsch, S, Stevenson, BJ, Flegel, V, Bucher, P, Jongeneel, CV, and Antonarakis, SE. (2002) Numerous Potentially Functional but Non-Genic Conserved Sequences on Human Chromosome 21. Nature, 420, 578-582.

[10] Bejerano, G, Pheasant, M, Makunin, I, Stephen, S, Kent, WJ, Mattick, JS, and Haussler, D. (2004) Ultraconserved Elements in the Human Genome. Science, 304, 1321-1325.

[11] Ahituv, N, Zhu, Y, Visel, A, Holt, A, Afzal, V, Pennacchio, LA, and Rubin, EM. (2007) Deletion of Ultraconserved Elements Yields Viable Mice. PLOS, 5, 1906-1911.

[12] Leabman,M.K., Huang,C.C., DeYoung,J., Carlson,E.J., Taylor,T.R., de la Cruz,M., Johns,S.J., Stryke,D., Kawamoto,M., Urban,T.J. et al. (2003) Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. Proc. Natl Acad. Sci. USA, 100, 5896–5901.

[13] Urban, T.J., Sebro, R., Hurowitz, E.H., Leabman, M.K., Badagnani, I., Lagpacan, L.L., Risch, N. and Giacomini, K.M. (2006) Functional genomics of membrane transporters in human populations. Genome Res., 16, 223–230.

[14] Stephens, J.C., Schneider, J.A., Tanguay, D.A., Choi, J.C., Acharya, T., Stanley, S.E., Jiang, R., Messer, C.J., Chew, A., Han, J., et al. 2001. Haplotype Variation and Linkage Disequilibrium in 313 Human Genes. *Science* **293**: 489-493.

[15] Hinds, D. A. 2005. The biology of genomes: sequence gives way to function. *Genome Biol.* **6**: 342.

[16] Myrick, A., Sarr, O., Dieng, T., Ndir, O., Mboup, S., and Wirth, D.F. 2005. Analysis of the genetic diversity of the Plasmodium Falciparum multidrug resistance gene 5' upstream region. *Am. J. Trop. Med. Hyg.* **72**: 182-188.

[17] Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C.R., Lim E.P., Kalyanaraman, N., Nemesh, J., et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Gen.* **22**: 231-238.

[18] Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., and Chakravarti, A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature* **22**: 239-247.

[19] Risch, N.J. 2000. Searching for genetic determinants in the new millennium. *Nature* **405**: 847-856.

[20] Langford, C.J., Klinz, F., Donath, C., and Gallwitz, D. 1984. Point Mutations Identify the Conserved, Intron-Contained TACTAAC Box as an Essential Splicing Signal Sequence in Yeast. *Cell* **36**: 645-653.

[21] Vijayraghavan, U., Parker, R., Tamm, J., Iimura, Y., Rossi, J., Abelson, J., and Guthrie, C. 1986. Mutations in conserved intron sequences affect multiple steps in the yeast splicing pathway, particularly assembly of the spliceosome. *The EMBO Journal* **5**: 1683-1695.

[22] Yu, L., Heere-Ress, H., Boucher, B., Defesche, J.C., Kastelein, J., Lavoie, Marc-Andre, and Genest, J. 1999. Familial hypercholesterolemia. Acceptor splice site (G→C) mutation in intron 7 of the LDL-R gene: alternate RNA editing causes exon 8 skipping or a premature stop codon in exon 8. *Atherosclerosis* **146**: 125-131.

[23] Webb, J.C., Patel, D.D., Shoulders, C.C., Knight, B.L., and Soutar, A.K. 1996. Genetic variation at a splicing branch point in intron 9 of the low density lipprotein (LDL)-receptor gene: a rare mutation that disrupts mRNA splicing in a patient with familiar hypercholesterolaemia and a common polymorphism. *Hum. Mol. Gen.* **5**: 1325-1331.

[24] Kuivenhoven, J.A., Weibusch, H., Pritchard, P.H., Funke, H., Benne, R., Assmann, G., and Kastelein, J.P. 1996. An Intronic Mutation in a Lariat Branchpoint Sequences Is a Direct Cause of an Inherited Human Disorder (Fish-Eye Disease). *J. Clin. Invest.* **98**: 358-364.

[25] Scott L.J., Mohlke K.L., Bonnycastle L.L., Willer C.J., Li Y., Duren W.L., et al. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**: 1341–1345.

[26] Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., et al. 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**: 870-874.

[27] Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D.P., Thompson, D., et al. 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**: 1087-1093.

[28] Freimer, N.B., and Sabatti, C. 2007. Variants in common diseases. *Nature* **445**: 828-823.

[29] Cartegni, L., Chew, S.L., and Krainer, A.R. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.* **3**: 285-298.

[30] Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., and Darnell, J.E. 2000. Molecular Cell Biology, 4th ed.

[31] Moore, M.J. 2002. Intron recognition comes of Age. *Nat. Struct. Biol.* **7**: 14-16.

[32] McGrail, J.C., and O'Keefe, R.T. 2008. The U1, U2 and U5 snRNAs crosslink to the 5' exon during yeast pre-mRNA splicing. *Nucleic Acids Res.* **36**: 814-825.

[33] Newman, A.J., Teigelkamp, S., and Beggs, J.D. 1995. snRNA interactions at 5' and 3' splice sites monitored by photoactivated crosslinking in yeast spliceosomes. *RNA* **1**: 968-980.

[34] Kramer, A. 1996. The Structure and Function of Proteins Involved in Mammalian pre-mRNA Splicing. *Annu. Rev. Biochem.* **65**: 367-409.

[35] Staley, J.P., and Guthrie, C. 1998. Mechanical Devices of the Spliceosome: Motors, Clocks, Springs, and Things. *Cell* **92**: 315-326.

[36] Newman, A.J. 1997. The Role of U5 snRNP in pre-mRNA Splicing. *EMBO Journal* **16**: 5797-5800.

[37] Wu, S., Romfo, C.M., Nilsen, T.W., and Green, M.R. 1999. Functional Recognition of the 3' Splice Site AG by the Splicing Factor U2AF35. *Nature* **402**: 832-835.

[38] Schneider, J.A., Pungliya, M.S., Choi, J.Y., Jiang, R., Sun, X.J., Salisbury, B.A., and Stephens, J.C. 2003. DNA variability of human genes. *Mech. of Ageing and Dev.* **124**: 17-25.

[39] Abril, J.F., Castelo, R., and Guigo, R. 2004. Comparison of splice sites in mammals and chicken. *Genome Res.* **15**: 111-119.

[40] Kol, G., Galit, L., and Ast, G. 2005. Human-mouse Comparative Analysis Reveals that Branch-site Plasticity Contributes to Splicing Regulation. *Hum. Mol. Genet.* **14**: 1559-1568.

[41] Ruskin, B., Krainer, A.R., Maniatis, T., and Green, M.R. 1984. Excision of an Intact Intron as a Novel Lariat Structure during pre-mRNA Splicing in Vitro. *Cell* **38**: 317-331.

[42] Zeitlin, S., and Efstratiadis, A. 1984. In Vivo Splicing Products of the Rabbit β-Globin Pre-mRNA. *Cell* **39**: 589-602.

[43] Ruskin, B., Greene, J.M., and Green, M.R. 1985. Cryptic Branch Point Activation Allows Accurate In Vitro Splicing of Human β-Globin Intron Mutants. *Cell* **41**: 833-844.

[44] Reed, R., and Maniatis, T. 1985. Intron Sequences Involved in Lariat Formation during Pre-mRNA Splicing. *Cell* **41**: 95-105.

[45] Padgett, R.A., Konarska, M.M., Aebi, M., Hornig, H., Weissmann, C., and Sharp, P.A. 1985. Nonconsensus branch-site sequences in the in vitro splicing of transcripts of mutant rabbit β-globin genes. *Proc. Natl. Acad. Sci.* **82**: 8349-8353.

[46] Hornig, H., Aebi, M., and Weissmann, C. 1986. Effect of mutations at the lariat branch acceptor site on β-Globin pre-mRNA splicing in vitro. *Nature* **324**: 589-591.

[47] Hartmuth, K., and Barta, A. 1988. Unusual Branch Point Selection in Processing of Human Growth Hormone Pre-mRNA. *Mol. and Cell. Biol.* **8**: 2011-2020.

[48] Gozani, O., Potashkin, J., and Reed, R. (1998). A Potential Role for U2AF-SAP 155 Interactions in Recruiting U2 snRNP to the Branch Site. Molecular and Cellular Biology 18, 4752-4760.

[49] Query, C.C., McCaw, P.S., and Sharp, P.P. (1997). A Minimal Spliceosomal Complex A Recognizes the Branch Site and Polypyrimidine Tract. Molecular and Cellular Biology 17, 2944-2953.

[50] Kralovicova, J., Houngninou-Molango, S., Kramer, A., and Vorechovsky, I. (2004). Branch Site Haplotyoes That Control Alternative Splicing. Human Molecular Genetics 13, 3189-3202.

[51] Khan, S.G., Metin, A., Gozukara, E., Inui, H., Shahlavi, T., Muniz-Medina, V., Baker, C.C., Ueda, T., Aiken, J.R., Schneider, T.D., et al. (2004). Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk. Human Molecular Genetics 13, 343-352.

[52] Hamlington, J.D., Clough, M.V., Dunston, J.A., and McIntosh, I. (2000). Deletion of a branch-point consensus sequence in the LMX1B gene causes exon skipping in a family with nail patella syndrome. European Journal of Human Genetics 8, 311-314.

[53] Vivenza D., Guazzarotti L., Godi M., Frasca D., di Natale B., Momigliano-Richiardi P., Bona G., Giordano M. (2006). A novel deletion in the GH1 gene including the IVS3 branch site responsible for autosomal dominant isolated growth hormone deficiency. Journal of Clinical Endocrinology Metabolism 91, 980-986.

[54] Vuillaumier-Barrot, S., Le Bizec, C., De Lonlay, P., Madinier-Chappat, N., Barnier, A., Dupré, T., Durand, G., and Seta, N. (2006). PMM2 intronic branch-site mutations in CDG-Ia. 87, 337-340.

[55] Ast, G., Pavelitz, T., Weiner, A.M. (2001). Sequences upstream of the branch site are required to form helix II between U2 and U6 snRNA in a trans-splicing reaction. 29, 1741-1749.

[56] Chamary, J.V., Parmley, J.L., and Hurst, L.D. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7**: 98-108.

[57] Nackley, A.G., Shabalina, S.A., Tchivileva, I.E., Satterfield, K., Korchynskyi, O., Makarov, S.S., Maixner, W., and Diatchenko, L. 2006. Human Catechol-O-Methyltransferase Haplotypes Modulate Protein Expression by Altering mRNA Secondary Structure. *Science* **314**: 1930-1933.

[58] Wang, J., Smith, P.J., Krainer, A.R., and Zhang, M.Q. 2005. Distribution of SR Protein Exonic Splicing Enhancer Motifs in Human Protein-Coding Genes. *Nucleic Acids Res.* **33**: 5053-5062.

[59] Kimchi-Sarfaty, C., Oh J.M., Kim, I.W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V., and Gottesman, M.M. (2007). A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science* **315**: 525-528.

[60] Lew, J.M., Fei, Y.L., Aleck, K., Blencowe, B.J., Weksberg, R., and Sadowski, P.D. (2004). CDKN1C Mutation in Wiedemann-Beckwith Syndrome Patients Reduce RNA Splicing Efficiency and Identifies a Splicing Enhancer. American Journal of Medical Genetics 127A, 268-276.

[61] Seth, P., Miller, H.B., Lasda, E.L., Pearson, J.L., and Garcia-Blanco, M.A. (2008). Identification of an Intronic Splicing Enhancer Essential for the Inclusion of FGFR2 Exon IIIc. The Journal of Biological Chemistry 283, 10058-10067.

[62] McCarthy, E.M.S., and Phillips, J.A. (1998). Characterization of an Intron Splice Enhancer that Regulates Alternative Splicing of Human GH pre-mRNA. Human Molecular Genetics 7, 1491-1496.

[63] Matsushima, M., Kobayashi, K., Emi, M., Saito, H., Saito, J., Suzumori, K., and Nakamura, Y. (1995). Mutation Analysis of the BRCA1 Gene in 76 Japanese Ovarian Cancer Patients: Four Germline Mutations, but No Evidence of Somatic Mutation. Human Molecular Genetics 4, 1953-1956.

[64] Matlin, A.J., Clark, F., and Smith, C.W.J. (2005). Understanding Alternative Splicing: Towards a Cellular Code. Nature 6, 386-398.

[65] Venables, J.P. (2007). Downstream Intronic Splicing Enhancers. Federation of European Biochemical Societies Letters 581, 4127-4131.

[66] Yeo, G., Hoon, S., Venkatesh, and Burge, C.B. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *PNAS* **101**: 15700-15705.

[67] Stryke, D., Huang, C.C., Kawamoto, M., Johns, S.J., Carlson, E.J., Deyoung, J.A., Leabman, M.K., Herskowitz, I., Giacomini, K.M., and Ferrin, T.E. (2003).

SNP Analysis and Presentation in the Pharmacogenetics of Membrane Transporters Project. Pac. Symposium on Biocomputing 8, 535-547.

[68] Crawford, D.C., Carlson, C.S., Rieder, M.J., Carrington, D.P., Yi, Q., Smith, J.D., Eberle, M.A., Kruglyak, L., and Nickerson, D.A. (2004). Haplotype Diversity across 100 Candidate Genes for Inflammation, Lipid Metabolism, and Blood Pressure Regulation in Two Populations. Am. J. Hum. Genet. 74, 610-622.

[69] Livingston, R.J., von Niederhausern, A., Jegga, A.G., Crawford, D.C., Carlson, C.S., Rieder, M.J., Gowrisankar, S., Aronow, B.J., Weiss, R.B., and Nickerson, D.A. (2004). Pattern of Sequence Variation Across 213 Environmental Response Genes. Gen. Res. 14, 1821-1831.

[70] Hartl, D.L., and Clark, A.G. (1997). Principles of Population Genetics 3rd ed, 37-69.

[71] Tajima, F. (1989). Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. Genetics 123, 585-595.

[72] Zhuang, Y, and Weiner, AM (1989) A Compensatory Base Change in Human U2 snRNA Can Suppress a Branch Site Mutation. Genes & Dev. 3, 1545-1552.

[73] Keller, EB, and Noon, WA. (1984) Intron Splicing: A Conserved Internal Signal in Introns of Animal Pre-mRNAs. PNAS, 81, 7417-7420.

[74] Lim, LP, and Burge, CB. (2001) A Computational Analysis of Sequence Features Involved in Recognition of Short Introns. PNAS, 98, 11193-11198.

[75] Harris, NL, and Senapathy, P. (1990) Distribution and consensus of branch point signals in eukaryotic genes: a computerized statistical analysis. Nucl. Acids Res. 18, 3015-3019.

[76] Wakeley, J. (1996) The Excess of Transitions Amongh Nucleotide Substitutions: New Methods of Estimating Transition Bias Underscore its Significance. TREE, 11, 158-162.

[77] Liu, S., W. Liu, J. L. Jakubczak, G. L. Erexson, K. R. Tindall, R. Chan, W. J. Muller, S. Adhya, S. Garges, and G. Merlino. 2002. Genetic instability favoring transversions associated with ErbB2-induced mammary tumorigenesis. *Proc. Natl. Acad. Sci. USA* 99: 3770–3775.

[78] Freudenberg-Hua, Y, Freudenberg, J, Kluck, N, Cichon, S, Propping, P, and Nöthen, MM. (2003) Single Nucleotide Variation Analysis in 65 Candidate Genes for CNS Disorders in a Representative Sample of the European Population. Genome Res., 13, 2271-2276.

[79] McCullough, A.J. and Berget, S.M. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol. Cell. Biol., 17, 4562-4571.

[80] McCarthy, EMS, and Phillips, JA. (1998) Characterization of an Intron Splice Enhancer that Regulates Alternative Splicing of Human GH pre-mRNA. Human Molecular Genetics, 7, 1491-1496.

[81] Coolidge, CJ, Seely, RJ, and Patton, JG. (1997) Functional analysis of the polypyrimidine tract in pre-mRNA splicing. Nucleic Acids Research, 25, 888-896.

[82] Burrows, NP, Nicholls, AC, Richards, AJ, Luccarini, C, Harrison, JB, Yates, JRW, and Pope, FM. (1998) A Point Mutation in an Intronic Branch Site Results in Aberrant Splicing of COL5A1 and in Ehlers-Danlos Syndrome Type II in Two British Families. Am. J. Hum. Genet., 63, 390-398.

[83] Janssen, RJ, Wevers, RA, Haussler, M, Luyten, JA, Steenbergen-Spanjers, GC, Hoffman, GF, Nagatsu, T, and Van Den Heuvel, LP. (2000) A branch site mutation leading to aberrant splicing of the human tyrosine hydroxylase gene in a child with a severe extrapyramidal movement disorder. Ann. Hum. Genet, 64, 375-382.

[84] Durbin, R, Eddy, S, Krogh, A, and Mitchison, G. (1998) Biological sequence analysis.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

DEC 8, 2009

Date