Spring 2022

Conservation and Prevalence of Sequence Paired Sites in Humans

Punit Sundar
San Jose State University

Conservation and Prevalence of Sequence Paired Sites in Humans

A Project

Presented to
The Faculty of the Department of Computer Science
San Jose State University

In Partial Fulfillment
Of the Requirements for the
Degree Master of Science in Bioinformatics

By
Punithavathi Sundaramurthy
May 2022

ABSTRACT

The completion of the Human Genome Project in 2003 made it possible to leverage the power of computers to quicken biological discoveries. The human genome contains rich information relevant to the proper functions of cells that is computationally arduous and costly to investigate in a laboratory setting. Specifically, gene expression levels in cells can be impacted when functionally relevant motifs in the DNA are mutated preventing proper transcription factor binding. To study the consequences of such mutations, it is necessary to first identify such functionally relevant regions of the genome using computational approaches. A signaling pathway that is of particular interest and important to cell differentiation, proliferation, apoptosis, and homeostasis is the Notch1 pathway. Mutations in the Notch1 pathway can lead to developmental defects, early embryonic death, and various forms of cancers such as T-cell lymphoblastic leukemia. The transcription machinery required for the expression of some Notch1 target genes include the binding of the transcription factor CSL to Notch response elements (NRE) located in the promoter region of Notch1 target genes and the formation of a transcription complex. Sequence paired sites (SPSs) are a pair of NREs with a varying spacer region that allow for the transcription factor CBF1 to bind to the cognate DNA which aids in the formation of a dimer regulating activation of downstream Notch1 target genes. Prior research has shown that mutations within the SPS have been found to significantly affect the activation of *Hes1*. This project will introduce a novel bioinformatics tool used to find potential SPSs in the genome and identify downstream genes for further study.

Index terms - Sequence Paired Site (SPS), Notch1, Notch Transcription Complex (NTC), Notch Response Element (NRE)

ACKNOWLEDGEMENTS

Throughout my time in the bioinformatics master's program, I have received unbelievable support from my project advisor Dr. Wendy Lee. She has played a big role in not only my master's project, but she has also helped in my professional development by giving me the opportunity to teach and publish my research. She is an inspiration to all of her students.

I would like to thank my committee members Dr. Brandon White and Dr. Philip Heller who have been a great resource as I developed my master's project. Working with all of my committee members, faculty in the CS department, and my peers in the bioinformatics program has been a rewarding experience.

I would also like to thank my parents and friends for their support. They have been a great source of comfort through the challenging times in grad school.

TABLE OF CONTENTS

List of Figures

List of Tables

1. INTRODUCTION

1.1 Biological Background

Notch receptors (Notch1, Notch2, Notch3, Notch4) are a group of transmembrane proteins involved in the regulation of Notch target genes which promote and inhibit cell fate decisions [1]. The extracellular part of Notch receptors consist of Epidermal growth factor (EGF)-like repeats which bind to an adjacent cell's ligands (Delta, Jagged, and Serrate) (Figure 1).

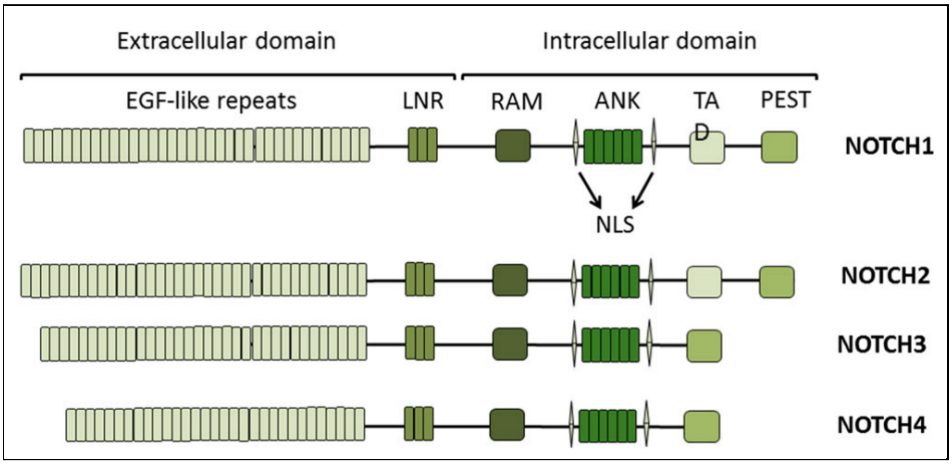


Figure 1. The four Notch receptors found within the Notch family of transmembrane proteins. EGF-like repeats are part of the extracellular components which bind to adjacent cell ligands. The intracellular domain translocates to the nucleus where it is involved in the regulation of Notch target genes. Adapted from Cell Commun. Signal. (9), copyright 2015.

This triggers a cascade of cleavages by ADAM metalloprotease followed by γ-secretase releasing the Notch intracellular domain (NICD) (Figure 2). The NICD then translocates to the nucleus where it binds to the transcription factor CSL (CBF1, Suppressor of Hairless, Lag-1). CSL acts as a transcriptional repressor but upon the binding of NICD, it recruits coactivators such as the Mastermind-like protein (MAM) to form the Notch transcription complex (NTC) enabling the upregulation of the downstream target gene [1]. Specifically, binding of MAM to the NICD induces dimerization which is required for the activation of some Notch target genes such as *Hes1* [2]. The site which CSL binds to in the cognate DNA is referred to as a Notch

response element (NRE) and prior research has shown the importance of the binding consensus sequence RTGRGAR (R = A/G) to promoter activation of Notch target genes [3].

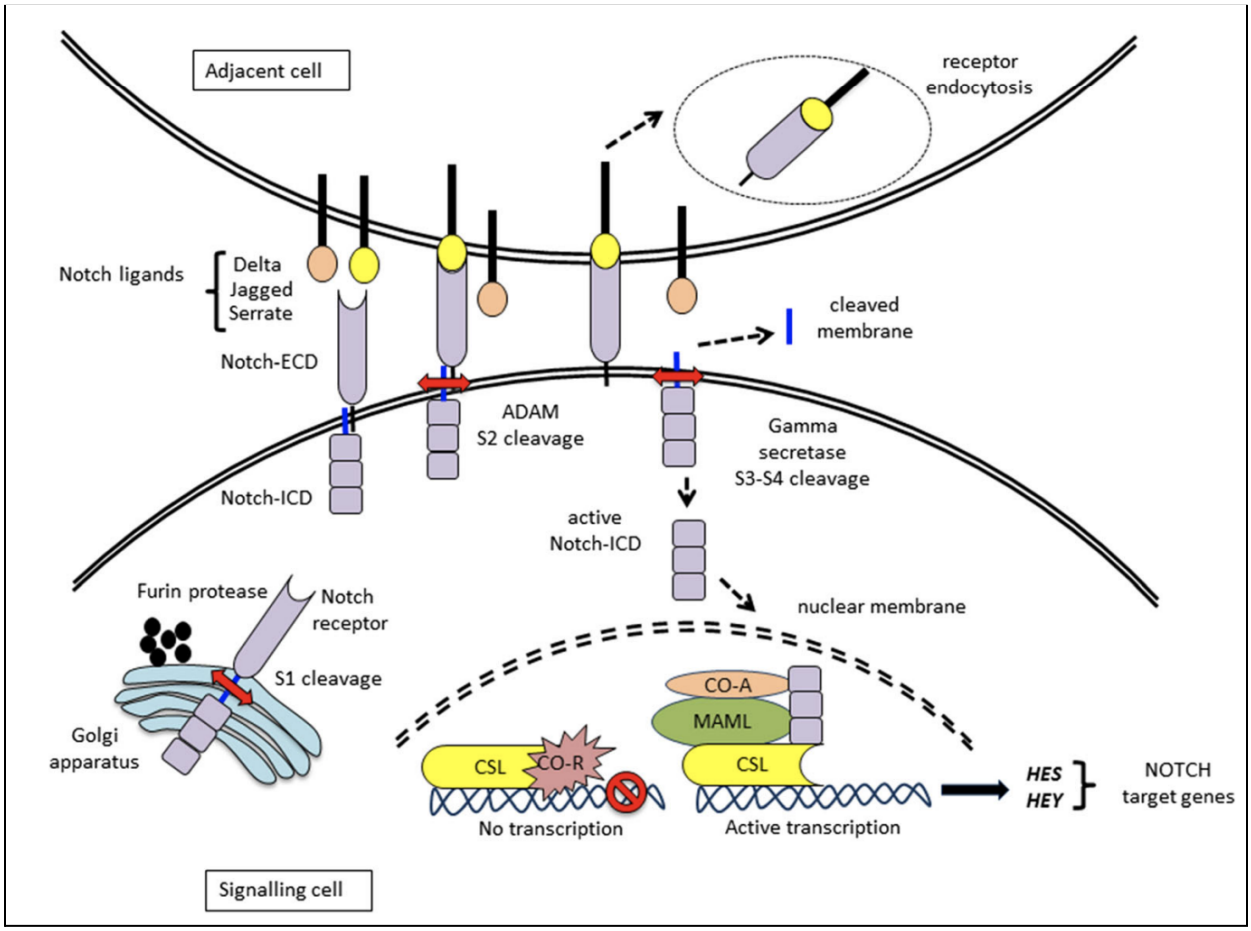


Figure 2. Notch signaling pathway.Adapted from Cell Commun. Signal. (9), copyright 2015.

Previously proposed models for Notch activation suggest the formation of dimers, or combinations of both monomers and dimers [2]. Specifically, two NREs separated by a spacer region between 15-17 base pairs in length have been thought to aid in the formation of the NTC dimer [4]. These two NREs are referred to as sequence-paired sites (SPS) and exist in the promoter region of Notch target genes (Figure 3) and up to millions of base pairs upstream in the case of enhancers, making computational extractions to study such sites challenging [5]. Using the well-studied *Hes1* SPS, the CSL binding site 1 for dimeric binding is 5'-GTGGGAA-3' and site 2 is the reverse complement (5'-TTCACACG-3') [6]. Studies have shown that site 2 can

contain “cryptic” sites that are not typically known to bind to CSL but do bind if there is strong affinity for the site 1 motif. An example of a cryptic site is site 2 of the *Hes5* SPS (5’-ACTCCAG-3’). Using both *Hes1* and *Hes5* site 2 motifs (5’-TTCACAC-3’ and 5’-ACTCCAG-3’ respectively), the consensus motif for site 2 can be deduced to WYYMCAS (W=A/T, Y=C/T, M=A/C, S=G/C) using the International Union of Pure and Applied Chemistry (IUPAC) conventions allowing for slight site 2 variability.

Exactly how each of the four Notch transmembrane proteins affect Notch signaling is not yet well understood. The NTC recruits other transcriptional coactivators such as p300, a histone acetyltransferase, and MAM recruits a cyclin dependent kinase (CDK) 8 [7]. Mutations within the nucleotide sequence of the SPS and surrounding transcriptional elements for proper Notch target gene activation have been shown to significantly modify activation levels of downstream Notch target genes such as in *Hes1* [8]. The orientation of the NREs is also important. The wild-type of the SPS follows a head and tail orientation (head = forward strand, tail = reverse strand). Some Notch target genes such as *Hes1* and *Hes5* also require a TATA box upstream of the gene transcription start site, and the spacing between the NREs and the TATA box can also impact activation levels [9].

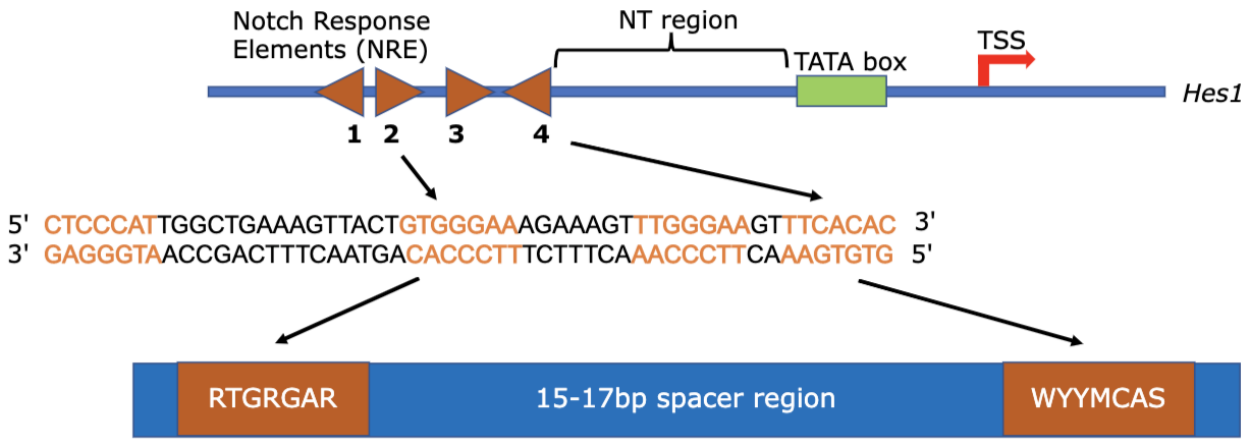


Figure 3. Four NREs and a TATA box upstream of *Hes1*. NREs 2 and 4 make up the SPS and have a spacer length of 16 base pairs. Spacer lengths can vary from 15-17bp for other Notch target genes. The TATA box is located 24 base pairs downstream of the SPS and 30 base pairs upstream of the transcription start site (TSS).

Apart from CSL, other transcription factors including Gal4 and TP53, bind to motifs with spacer regions aiding in dimer formation. Gal4 in *Saccharomyces cerevisiae* binds to the consensus sequence CGG[N*11]CCG (N=A/T/C/G) with a spacer length of 11 nucleotides and is important to fungal regulatory processes such as galactose metabolism and antifungal drug resistance [10]. The Gal4 ortholog in *Candida Albicans* binds to the same motif, but is not involved in galactose metabolism highlighting how the same motifs can serve different functions in related species. TP53 follows the motif 5'-RRRCWWGYYY[N*0-21]RRRCWWGYYY-3' with a spacer length between 0 to 21 nucleotides [11]. Studying transcription factor binding sites (TFBS) across organisms and their effects on regulation of target gene expression can help elucidate the molecular mechanisms required for transcriptional activation.

Many different approaches can be used to identify the required regulatory elements such as TFBS to study how they affect downstream target genes. This includes reporter gene assays, high-throughput methods, and computational approaches [12]. Plasmids with a motif of interest and a downstream target gene can be used to test how mutations in the motif might affect gene expression based on a defined measurement such as fluorescence. However, this approach can be

expensive, and the TFBS has to be known prior to being placed appropriately within the plasmids. High-throughput methods such as chromatin immunoprecipitation (ChIP) experiments are used to identify the regions of the genome that are bound to a protein of interest *in vivo*. An advantage of this method is that it can be used to identify enhancers that are not proximal to the promoter region. One disadvantage, however, is that an antibody specific to the transcription factor has to be created for ChIP experiments to obtain the precipitate (bound protein and genomic sequence). In addition, some transcription factors might require the binding of additional cofactors to enable proper gene expression. Analyzing ChIP experiment data can also pose problems as software tools are required, and this can limit a researcher who is not accustomed to building or running bioinformatics pipelines. Computational approaches to identify TFBS can be used to narrow down potential motif targets for further study in a laboratory setting through reporter gene assays or other methods. One disadvantage of this approach is the lack of tools available to identify complex TFBS and the downstream target gene for motifs that contain varying spacer lengths.

1.2 Current Approaches for Motif Identification

Current computational tools focus on identifying motifs in the genome that could potentially be necessary for transcription factor binding, but tools flexible enough for users to search for their own complex motifs of interest containing varying spacer lengths and downstream gene information do not currently exist.

1.2.1 Position Weight Matrices

One of the most common approaches used to identify motifs in a given sequence is a position weight matrix (PWM). A PWM is an array of values based on a motif training set of sequences of the same length that represents the frequency of bases at each position in the

sequence [13]. If a nucleotide does not exist at a specific position in the sequence, that nucleotide has a frequency of zero. However, genomic sequence variations can exist, such as point mutations, and to account for this variability, pseudo-counts are used to obtain a frequency slightly higher than zero. Once a PWM is created, it can be used to scan a long sequence, such as the entire human genome, to identify potential motifs that are similar to the training set. If a sequence during the scan meets a threshold value, that sequence is extracted as a potential motif. This works well for monomeric motifs which do not contain varying spacer region lengths.

PWMs, however, calculate probabilities of each nucleotide in a sequence independently. A motif known to always have three specific nucleotides next to each other, for example, will not be taken into consideration when a PWM is created.

1.2.1.1 FIMO

A tool within the MEME Suite called Find Individual Motif Occurrences (FIMO), the user is able to upload motifs and the database the user would like to search on a web interface (Figure 4) [14]. A command line version of the tool is also available for installation. An array of options for databases are provided such as the UCSC mammal genomes and specific genome versions. FIMO accepts user motifs in a meme-formatted file or typed motifs that are automatically turned into a PWM. IUPAC conventions can be used, but the motifs have to be of the same length. If multiple motifs of different lengths are given as input, FIMO automatically converts all of the input motifs into a single PWM which changes the original nucleotide frequency. For example, if the following two motifs are given as input, RTGRGAR[N*16]HYYHCAS and RTGRGAR[N*17]HYYHCAS, such that the second motif contains one extra N (N=A/T/C/G) to account for a varying spacer length, FIMO converts the motifs into the following monomeric motif RTGRGAR[N*17]HYYYMV. This combined motif

has a slightly different nucleotide composition than the original input motifs which could potentially extract false positive motifs. In addition to the spacer length disadvantage, FIMO search can be time consuming as the program searches the entire genome unless the user uploads their own custom database. This process has to be repeated multiple times if the user wants to search for multiple motifs containing varying spacer lengths which linearly increases time if computing resources are not available for multiple motif submissions. Though some disadvantages exist, FIMO can still be used to search for one motif of interest at a time and can provide motif matches for a full genome.

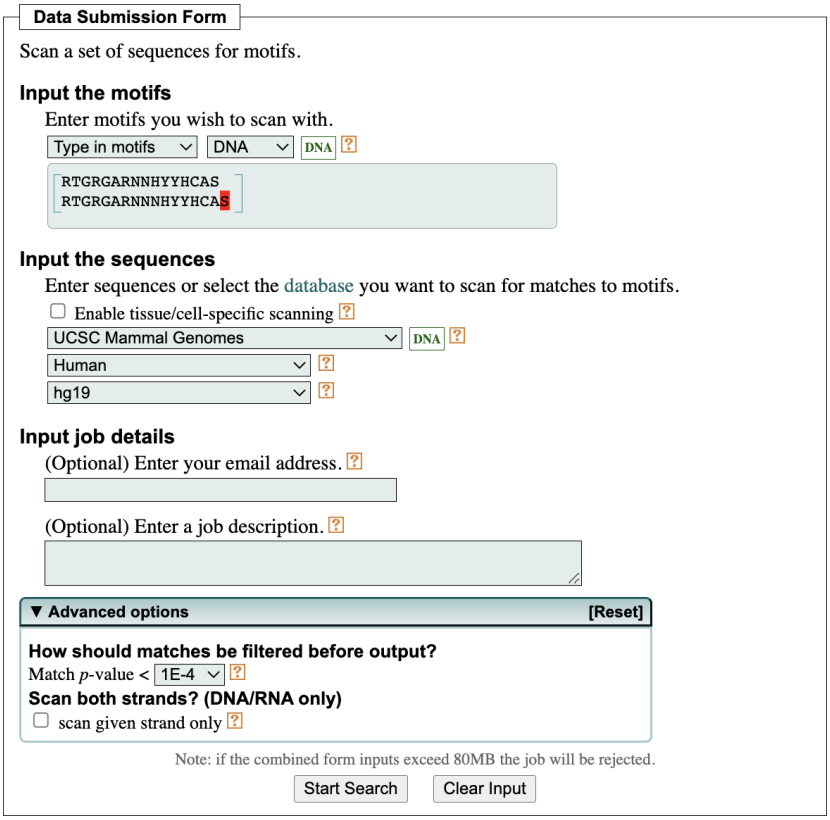


Figure 4. FIMO's web interface.

1.2.2 Profile Hidden Markov Models

An approach to identify motifs that takes into account nucleotide dependencies unlike PWMs is a Hidden Markov Model (HMM). In the context of biological sequence motifs, an

HMM is a probabilistic model used to optimize emission and transition probabilities from sequence training data to allow for motif detection against a sequence database [15]. An HMM requires a set of N states, transition probabilities, a set of observations, emission probabilities, initial probability distribution of all states, and assumes that the probability of a state only depends on the previous state [16]. HMMs are extensively used to classify protein sequences into their respective protein families [15].

The example shown in Figure 5 illustrates a profile HMM model for a multiple sequence alignment (MSA) of amino acid sequences which can be used to train an HMM to classify whether a sequence belongs to a specific protein family or not [17]. MSAs are used to align more than two sequences that allow for the discovery of important motifs or domains that otherwise might not be possible to obtain using pairwise alignment, sometimes creating gaps in the process [18]. Homologous sequences might contain conserved regions that are important for transcription factor binding or can be used to predict the structure of proteins. The states in the HMM diagram include the various matches, deletions, and insertions. Transition probabilities represent the probability of moving from one state to another in a sequence and are denoted with the arrows. A set of observations are the individual amino acids that make up a sequence. Emission probabilities are the probability of a particular state generating an observation. The initial probability is the probability of starting a sequence with a particular amino acid in the initial state.

Unlike PWMs, HMMs are position dependent. HMMs can thus be more useful to detect sequences that contain strongly conserved motifs occurring consecutively compared to PWMs.

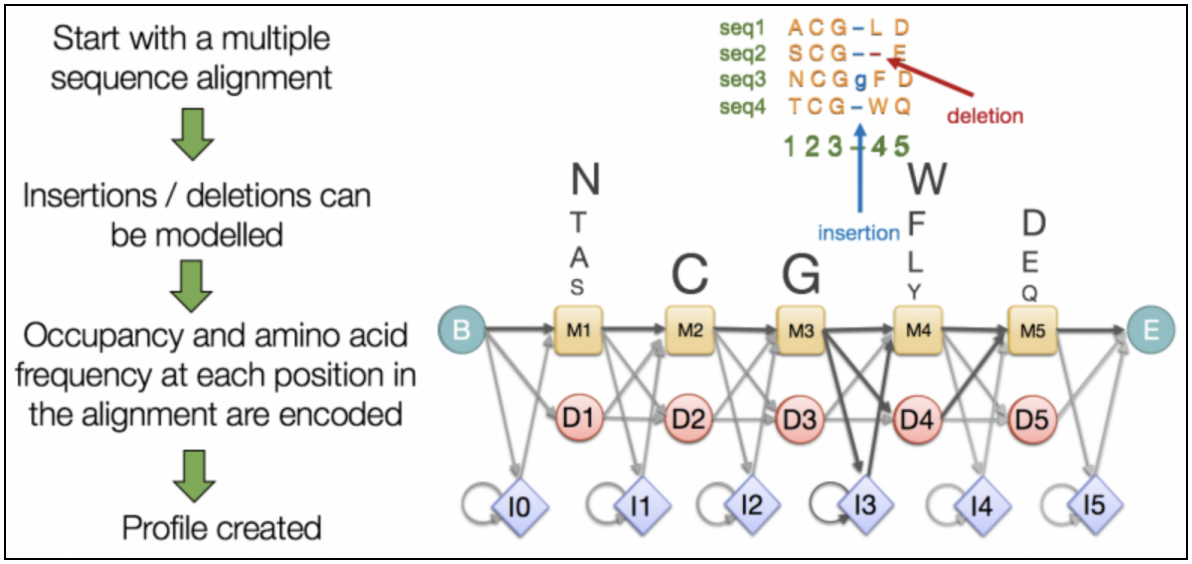


Figure 5. HMM to classify protein sequences. Gaps and insertions are modeled using the deletions (D1-D5) and insertions (I0-I5) states. Adapted from [17].

Though HMMs are useful to identify monomeric motifs if trained on a robust set of training data, it is not suited for motifs containing long spacer regions that are not conserved. For example, complex binding sites required for the formation of a dimer will negatively affect the emission and transition probabilities as the spacer region acts as noise. The spacers will also require building multiple HMMs, one for each length variation. HMMs also need to be optimized using training data, and this could be difficult to produce if a particular transcription factor is not well-studied.

1.2.2.1 HMMER

The HMM-based tool, hmmer, is used to identify statistically significant motifs given a sequence database [19]-[20]. It was originally used to classify sequences to their respective protein families. The command *hmmbuild* is used to train an HMM using a training dataset given in FASTA format. However, the training set sequences have to be of the same length. Once the HMM is built, the *nhmmer* command is used to identify potential motif matches in a large DNA

sequence file such as a chromosome FASTA file. The results will show the top hits which pass various hmmer filters.

1.2.3 Protein-Binding Microarray

Besides computational approaches to study transcription factor binding, high throughput methods such as protein binding microarrays (PBM) can be used. Microarrays are used to identify the binding sites of a transcription factor of interest. Custom designed microarrays exist to test which of all possible N-mer sequence variants can bind a specific transcription factor [21].

Bianco et. al. performed a PBM experiment targeting the NTC containing NICD, CSL, and MAM proteins [22]. The PBM experiment was used to identify potential monomeric binding sites of the NTC, and these sequences were ranked from -0.5 to +0.5 referred to as an enrichment score (E-score) using the Universal PBM Analysis Suite, the Seed-and-Wobble algorithm, and a modified form of the Wilcoxon-Mann-Whitney statistic. The closer to +0.5 the score, the higher the binding preference of the NTC to the sequence.

Using the monomeric data, artificial SPSs can be created. The product of the E-scores from the PBM data and fluorescence resonance energy transfer (FRET)-based assay signals were found to correlate ($R^2 = 0.80$) [23], suggesting a new method to extract SPSs from a sequence database. The FRET-based assay was used to identify the DNA sequences which allow for NTC dimerization. The E-score for site 1 of a monomeric sequence in the PBM data can be multiplied with another monomeric sequence used as site 2 to create an artificial SPS. This product score can be used to discriminate between potential SPSs and those sequences less likely to be SPSs. However, the data contains approximately 50,000 sequences and thus 2,500,000,000 combinations of artificial dimeric sequences need to be created. In addition, for each artificial dimeric sequence, more sequences can be created to account for varying spacer lengths. For

example, a dimeric sequence that can have 15, 16, and 17 base pair spacer lengths, would increase the number of sequences from 2,500,000,000 to 2,500,000,000*3 (or 7,500,000,000 artificial dimeric sequences). Currently existing computational tools, such as FIMO, are not efficient enough to parse through thousands of unique motifs in a time efficient manner.

1.2.4 Downstream Analyses of Sequence Data

Downstream analyses of sequence data can be used to make informative inferences of the results obtained from computational tools. For motifs, some of these downstream analyses can include MSA to cluster sequences to identify conserved regions not readily observable by sequence gazing and the creation of web logos used to identify the consensus sequence of various motifs. In the context of DNA sequences, clustering is based on building phylogenetic trees made of nodes and branches. The nodes are used to separate the sequences (branches) based on nucleotide differences between them [24]. They can be used to identify distinct features that separate the various nodes from one another including but not limited to TFBS motifs. To test the consistency of a phylogenetic tree, the bootstrap method can be used [25]. The bootstrap method resamples sequences in a dataset to generate a phylogenetic tree many times, and the final consensus phylogenetic tree built will contain bootstrap values that indicate the proportion of trees containing each consensus clade in the tree. Clustering can be used to cluster sequences that are highly similar to each other based on specific functional features such as TFBS. A method to visualize the consensus sequence data motif is a web logo [26]. Web logos are a visual representation of the frequency of nucleotides at each position in sequence data and help to quickly identify potentially conserved regions.

Extraction of potential SPSs from the human genome is needed to identify other genes besides well-studied genes like Hes1 and Hes5 that could be affected by mutations within the

SPS. No current bioinformatics tools report important information regarding the extracted motifs of interest including the downstream gene and TATA box. Manually identifying SPSs upstream of all human genes is time consuming and can introduce human errors. To extract SPSs, currently existing software tools, FIMO and hmmer, will be tested. A novel tool, XtractXact, was developed for this project to provide a new way to extract user motifs and provides additional gene and TATA box information not provided by FIMO or hmmer.

Using computational tools can quicken motif finding, but any results obtained by computational methods need to be further examined in the laboratory. To validate computational results of potential SPSs, scientists should perform molecular experiments such as ChIP-seq or gene assays to test whether NTC binding and dimerization occurs at those sites. Otherwise, results would not be void of false positives.

2. METHODS

To test whether the various existing computational approaches used to identify motifs can be used to extract potential SPSs from the human genome, FIMO and hmmer tools were tested. Due to some disadvantages of using FIMO and hmmer for this project, the XtractXact tool was created to specifically extract exact SPS motifs. Once potential SPSs are extracted, a bioinformatics pipeline can be built to extract downstream gene information to allow for further data exploration.

2.1 Current Computational Tools

To extract potential SPSs from the human genome, the site 1 motif RTGRGAR consensus is used as it is highly conserved compared to site 2 which is more variable (Table 1). The site 2 motifs used are HYYHCAS, WYYMCAS, and YTCHCAY. Human *Hes1* and *Hes5* SPS site 2 motifs were used to obtain the consensus WYYMCAS motif though this motif might vary in

other species (Table 1). The YTCHCAY motif is obtained from taking the reverse complement of the RTGRGAR motif and includes the SPS site 2 motif of the Xenopus species which changes the middle R in RTGRGAR from Y (Y=C/T) to an H (H=A/C/T). The HYYHCAS motif is used to broaden the search to include nucleotide variability. The following motifs will be extracted from the human and mouse genomes.

1. RTGRGAR[N*15-17bp]HYYHCAS
2. RTGRGAR[N*15-17bp]WYYMCAS
3. RTGRGAR[N*15-17bp]YTCHCAY

Human Hes1	T	T	C	A	C	A	C
Mouse Hes1	T	T	C	A	C	A	C
Rat Hes1	T	T	C	A	C	A	C
Zebrafish Hes1	T	T	C	A	C	A	C
Human Hes5	A	C	T	C	C	A	G
Mouse Hes5	A	C	C	C	C	A	G
Rat Hes5	A	C	C	C	C	A	G
Gallus Hes5	G	A	C	C	C	A	T
Xenopus Hes5	G	A	C	T	C	A	T
No Gallus & Xenopus	T/A	T/C	T/C	A/C	C	A	C/G
CONSENSUS	**W**	**Y**	**Y**	**M**	**C**	**A**	**S**

Table 1. Site 2 motifs across various species and the consensus.

2.1.1. FIMO

To use FIMO for motif identification, a random dimeric motif, DCYWSYS[N*16]MNKSGDA, was generated and was used as a test to show the disadvantages of using PWMs for exact motif extraction. The parameters listed in Table 2 were used to run FIMO. FIMO calculates the p-value and q-value for each motif occurrence which is based on a log-likelihood ratio [14]. Calculation of the q-value can increase motif finding time, so the --no-qvalue parameter was used. FIMO creates multiple output files from a tab-separated value (TSV) file to an html file to view on a web browser. The --text parameter is used to only allow

for a TSV file result to increase speed of FIMO. The --parse-genomic-coord allows the user to use UCSC style genomic coordinates for sequences in FASTA format. This will give results which not only include the motif match, but also its corresponding UCSC genomic coordinates. The database used to test FIMO include various promoter regions (2500, 1500, 500 base pairs upstream) of all genes in the human genome version 19 (hg19). In addition to the promoter region, the intragenic region of all human genes was also used.

Parameters
--no-qvalue
--text
--parse-genomic-coord

Table 2. Parameters used to run FIMO.

2.1.2 HMMER

To use hmmer, the SPS motif RTGRGAR[N*16]HYYHCAS was tested on chromosome 1. The human Hes5 SPS site 1 motif is located on chromosome 1, so hmmer was expected to return hits matching this motif. However, this returned no significant hits which could have been due to hmmer not being suitable for motifs with long unconserved spacer region [27]. Next, the RTGRGAR site 1 motif was used and this also did not return any hits on chromosome 1. Finally, variations of the RTGRGAR motif were used to create a training set of sequences in FASTA format to build an HMM as shown in Figure 6. The following command was used to build the HMM: hmmbuild --dna first_site.hmm motif1.txt. The HMM built using the FASTA file sequences is stored in the first_site.hmm file. To extract potential site 1 matches from the chromosome 1 database, the nhmmer command was used. However, the initial run of the nhmmer command with all default values resulted in zero hits once again, so parameters were modified. The following nhmmer command was used to extract potential matches in

chromosome 1: nhmmer --max --dna --w_length 7 -E 10 first_site.hmm chr1.fa > test_motif1.txt. The --max parameter was used to turn off all filtering algorithms to ensure the maximum number of hits returned in the output file. The --w_length parameter is the expected length of the hit. The value 7 was given since the length of the site 1 motif is 7 nucleotides in length without any gaps. The -E parameter is the threshold value needed for sequences to be reported in the results. The -E value of 10 was chosen so that a high number of hits would be reported.

The various filters hmmer uses include the SSV, bias, viterbi, and forward filters. The null model creates a log-odds bit score for sequences that hypothesizes non-homology. The SSV filter compares the query and the target by scanning for ungapped segments which have high log-odds scores. The bias filter is used to minimize any biases in the nucleotide composition of sequences that do not match the null model. The Viterbi filter uses the Viterbi algorithm to align the target sequences in the database with the HMM profile to find the optimal alignment. If the target meets a certain threshold value, it passes to the next filter. The forward filter calculates the likelihood of the target sequence against the HMM profile, and if the target sequence passes a certain threshold value, it is considered to be a likely significant hit.

```
>seq1
ATGAGAA
>seq2
ATGAGAG
>seq3
ATGGGAA
>seq4
ATGGGAG
>seq5
GTGGGAG
>seq6
GTGAGAA
```

Figure 6. Short training sequences to build HMM.

2.2 XtractXact

XtractXact was developed for this project to allow users to search for complex motifs of interest within promoter and intragenic regions of genes of a given genome quickly to retrieve exact motif matches for motifs with varying spacer lengths. This tool also addresses some of the shortcomings of PWMs and HMMs. The underlying method XtractXact uses is to search for site 1 in the promoter region of all genes or intragenic regions, and if site 1 matches exist, XtractXact moves over a spacer length and searches for the site 2 motif. If both site 1 and site 2 motifs exist, the match gets reported in a results file. XtractXact also reports the gene that is located downstream of the potential matched motifs and a potential TATA box. The TATA box is defined as the TATA[A/T]{1}A regular expression where TATA is followed by one occurrence of either an A or T nucleotide followed by A. The results file includes the motif matches in the genome, chromosomal locations of the motifs, whether the motif is on the positive or negative strand of the DNA, the gene located directly downstream of the motifs, the distance between the motifs and a potential TATA box, the distance between a potential TATA box and the transcription start site of the gene, the distance between the motif and the transcription start site of the gene, and other relevant information regarding the search criteria used.

Since XtractXact is currently used to search promoter and intragenic regions only, the promoter and intragenic regions along with gene locations need to be provided as input for running the tool. The UCSC Genome Browser contains sequence and annotation data for various genomes that can be used to create these required input files for XtractXact. The data required include chromosome FASTA files (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/chromosomes/) and the NCBI RefSeq genes gtf file (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/genes/). The chromosome

files contain the entire DNA for a given species and genome version. The NCBI RefSeq genes gtf file includes information about the various genes and their locations in the chromosomes. NCBI RefSeq is a curated database that is frequently updated to reflect current sequencing data.

With the input files, the run_gtf2gene.sh script is used to create TSV files required for XtractXact (Table 3-4). This step is only run once. The configuration file needed for the run_gtf2gene.sh script includes the URL for the human or mouse genome of interest found on the UCSC Genome Browser download section, the promoter length, and whether the run_gtf2gene.sh should extract the promoter or the intragenic regions of all genes. The required XtractXact files include multiple transcripts for each gene and hence XtractXact reports the same SPS but for different transcripts. To eliminate this redundancy, the initial XtractXact files were further modified to only contain the earliest and farthest genomic position of each gene. The entire genomic sequence (including introns and exons) are included in the initial XtractXact files.

```
# Get config file variables
# Make sure variables names are unique!

config_file=$1
source $config_file

# Make directories to store intermediary & result files
mkdir gtf_download
mkdir gene_transcript
mkdir motif_results

# Obtain gene gtf file
cwd=$(pwd)
wget -P ${cwd}/gtf_download/ ${GTF_LOCATION}${GTF_FILENAME}
gunzip ${cwd}/gtf_download/${GTF_FILENAME}
GTF_FILENAME="${GTF_FILENAME%.*}"

# Extract necessary columns from gtf file
awk '$3 == "transcript"  {print $1,$4,$5,$14,$7,$12}' ${cwd}/gtf_download/$
{GTF_FILENAME} | sed 's/"//g' | sed 's/;//g' | sed 's/ /\t/g' | sed 's/$/\t0/' > ${cwd}/gtf_download/
${SPECIES}_gtf.bed

# Remove alternative chromosomal locations & duplicates
python ${cwd}/scripts/remove_alternatives.py -b ${cwd}/gtf_download/${SPECIES}_gtf.bed -o
${cwd}/gene_transcript/${OUTPUT_GENE_FILENAME}_wo_alt
awk '!/chrMT/' ${cwd}/gene_transcript/${OUTPUT_GENE_FILENAME}_wo_alt | awk '!
seen[$1,$2,$3,$4]++'  > ${cwd}/gene_transcript/${SPECIES}_wo_dup

echo Complete...

# Run either promoter or intragenic region extraction
# Extracting N_bp upstream takes ~1 minute
echo starting extraction...
if [[ "$REGION" == "PROMOTER" ]]
then
   date
   python ${cwd}/scripts/front_extract.py -c ${CHR_DIR} -g ${cwd}/gene_transcript/$
{SPECIES}_wo_dup -bp ${N_BP} -o ${cwd}/gene_transcript/${SPECIES}_extracted_$
{N_BP}.txt
   date
else
   date
   python ${cwd}/scripts/gene_extract.py -c ${CHR_DIR} -g ${cwd}/gene_transcript/$
{SPECIES}_wo_dup -o ${cwd}/gene_transcript/${SPECIES}_intragenic_extract.txt
   date
fi
echo extraction complete…
```

Figure 7. The run_gtf2gene.sh script used to create the necessary input files for XtractXact.

```
# config file to retrieve GTF file to produce gene transcript file
# Author: Punit Sundar
# 11/18/21

# Comment out variables for mouse genome if extracting human and vice versa
# URL used to use wget

#GTF_LOCATION=https://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/genes/
#GTF_FILENAME=mm10.ncbiRefSeq.gtf.gz
#SPECIES=mouse
#OUTPUT_GENE_FILENAME=mouse_gene
#CHR_DIR=/home/psundar/data/genomes/mouse/mm10

GTF_LOCATION=https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/genes/
GTF_FILENAME=hg19.ncbiRefSeq.gtf.gz
SPECIES=human
OUTPUT_GENE_FILENAME=human_gene
CHR_DIR=/home/psundar/data/genomes/human/hg19

# Number of base pairs upstream to scan
N_BP=100

# Promoter or intragenic region?
REGION=PROMOTER
#REGION=INTRAGENIC
```

Figure 8. Config file required to run the run_gtf2gene.sh script. Variables can be changed as needed. The N_BP variable can be changed to the requested upstream region of genes and will not affect the run_gtf2gene.sh script when extracting intragenic regions.

Chromosome	Gene	Strand	Gene Start	Gene Stop	Extracted Start	Extracted Sequence	Accession ID
chrY	DDX11L16	-	59358329	59360854	59360954	ATGACTGCGCAAATTTGCCGGATTTCCTTTGCTGTTCCTGCATG TAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTT TCTTTTTGTTAA	NR_110561 .1
chrY	WASH6P	+	59354328	59358341	59354228	CAAGCAGGTGACCCTGACTTCAGAGCCCTTGCCTGAGGGCCT GGCCTGGCAGCTCTGCTGTTAGAAGCAGGAGGTGTGCAGGGG GTGGGGAGCAGCCCAG	WASH6P
chrY	WASIR1	-	59347235	59349501	59349601	TGATTTTCAGAGACTCCCCATGGCTGCCGCAAGGGGGCGCATG GCCCCTGGCAACTCGGGGCGCCGTGCACGCACCGCGTCTCAT CCACACAGCGGCAGT	NR_13804 8.1
chrY	IL9R	+	59330366	59343488	59330266	GCGCTTGTGGTTTCAGATGTGGGCGGCCTGTGTGAACCTGTCG TGCAAAGCTCACGTCACCAACTGCTGCAGTTATCTCCTGAATC AGGCTGAGGGTCTT	NM_00218 6.3
chrY	TRPC6P	-	59318025	59318922	59319022	TGAGGACTTAGAAGGAGCCTGGGCAAGTAAAGTCATCTTGAA GCTTCAGTTTCATGGTTAACCTTCCTCTGCTAGAGAAAAACAG ATTTAATGAGCGCAG	TRPC6P

Table 3. An example of the tab-separated promoter file created by the run_gtf2gene.sh script. File includes chromosomal locations of genes, extracted upstream sequence (100bp), and the transcript accession ID. The run_gtf2gene.sh script can be run multiple times to acquire additional upstream files other than 100bp such as 500bp, 1500bp, etc.

Chromosome	Gene	Strand	Gene Start	Gene Stop	Gene Sequence	Accession ID
chrY	DDX11L16	-	59358329	59360854	CTTGCCGTCAGCCTTTT… GCACACTGTTGGTTTCTGCTC	NR_110561.1
chrY	WASH6P	+	59354328	59358341	CCTCTGTGATCTTCTCC… AAAGAACTGAGCAGAAACCAA	WASH6P

Table 4. An example of the tab-separated intragenic file created by the run_gtf2gene.sh script. File includes chromosomal locations of genes, entire gene sequence, and the transcript accession.

Once the promoter and intragenic files are created, XtractXact can be used. XtractXact requires input motifs which are given in the form of text files. Currently, motifs are formatted such that each motif has its own row in the file (Figure 9). Motifs can follow IUPAC conventions allowing the user flexibility in their search.

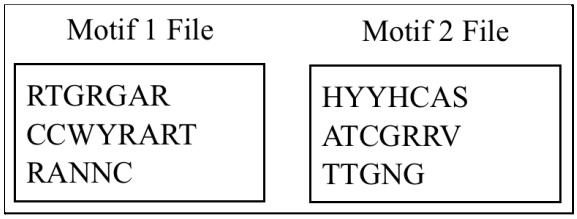


Figure 9. An example of the motif files needed for XtractXact. Each motif in file 1 are paired with each of the motifs in file 2 separated by the various spacer lengths requested.

In addition to motif files, various other criteria can be requested (Table 5). The main Python script (XtractXact.py) used to run the XtractXact tool requires multiple parameters. These include the species name (“human” or “mouse”), genome version (“hg19” or “mm10”), DNA strand (“-” or “+” or “both”), chromosome number (specific number of chromosomes or all chromosomes can be selected), one file with first motif sites, second file with second motif sites, spacer length (list of all spacers needed), and the promoter or intragenic region can be requested for the search. Other genome versions (“hg38”, “mm39”, etc.) can also be used if the chromosome FASTA files location is added to the XtractXact.py script’s “get_genome_path” function, and the variables within the config file (GTF_LOCATION, GTF_FILENAME, SPECIES, OUTPUT_GENE_FILENAME, and CHR_DIR) are modified. The XtractXact.py

script takes these parameters and identifies all motif matches separated by the various spacer lengths in the chromosome files corresponding to the promoter or intragenic regions, potential TATA boxes, and other useful information. The results of the motifs search are written to a TSV file in alphabetical order by gene name.

Parameters	**Example**
Genome	Human
Version	hg19
Strand	both
Chromosomes	chr1,chr2,chr5,chr8
Spacer	15,16,17
Gene	Promoter

Table 5. Various parameter options in XtractXact. Strand option includes minus, plus, and both strands. Gene option is used to search either the promoter or intragenic region.

XtractXact was used to obtain potential SPSs for the human (hg19) and mouse (mm10) genomes using the parameters shown in Table 6.

Parameters	**Human values**	**Mouse values**	**Description of Parameters**
--genome_species	human	mouse	Organism of choice
--genome_version	hg19	mm10	Genome version of organism
--strand	both	both	DNA strand to search
--chromosome	all	all	Chromosomes to search
--file1	input1.txt	input1.txt	File containing all site 1 motifs
--file2	input2.txt	input2.txt	File containing all site 2 motifs
--spacer	15,16,17	15,16,17	Spacer lengths between site 1 & 2
--gene	True/False	True/False	Promoter search (False) / Intragenic search (True)

Table 6. All parameters used in XtractXact to search SPS motifs in the human and mouse genome. The input1.txt file contains the single motif RTGRGAR. The input2.txt file contains the motifs HYYHCAS, WYYMCAS, and YTCHCAY.

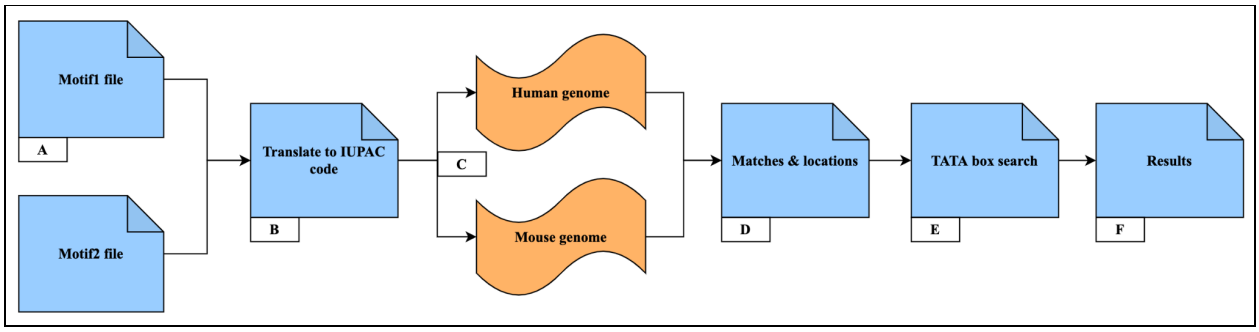


Figure 10. Complete XtractXact pipeline. A. Input motif files. B. Motifs translated to regular expressions from IUPAC code. C. The human or mouse genome is searched for the motifs. D. Motifs and locations are extracted from the genome. E. A search for potential TATA boxes upstream of the promoter region or within the genes. F. Tab-separated results file for motif matches is created for the user to explore.

To check whether the XtractXact results can be trusted, the known SPS sites for *Hes1* and *Hes5* in hg19 are searched in the XtractXact results (Table 7). The *Hes1* SPS occurs in the positive strand and the *Hes5* SPS occurs in the negative strand of the DNA, and this can test whether XtractXact processes both DNA strands correctly.

Gene	Strand	Spacer Region	Start	Stop	SPS Motif
Hes1	+	16	193853852	193853881	**GTGGGAA**AGAAAGTTTGGGAAGT**TTCACAC**
Hes5	-	16	2461754	2461783	**GTGGGAA**CGGCCGCGGCGCCCGG**ACTCCAG**

Table 7. Known SPS motifs used for XtractXact validation. The spacer region is the number of nucleotides between the first CSL binding site and the second CSL binding site. The bolded font represents site 1 and site 2 motifs of the SPS.

XtractXact was used to extract the three SPS motifs (RTGRGAR[N*15-17]HYYHCAS, RTGRGAR[N*15-17]WYYMCAS, RTGRGAR[N*15-17]YTCHCAY). Once motifs search results are obtained, the following steps are performed for downstream analyses including clustering and web logos.

1. split_results.sh - splits TSV results by spacer length
2. get_fasta.py - creates FASTA formatted files for all split TSV results

3. Cluster motif matches using Molecular Evolutionary Genetics Analysis (MEGA) X with the unweighted pair group method with arithmetic mean (UPGMA) algorithm for the HYYHCAS motif for human promoter results with spacer length 16 (Figure 11). The test phylogeny method included “None” or “Bootstrap method”. The bootstrap method was not used due to the sequences containing a large spacer region. This large spacer region can result in different clades during the resampling process as only the site 1 and site 2 motifs are conserved. All other parameters are set to default [28]. Within the substitution model, the maximum composite likelihood parameter is estimated by maximizing the sum of log-likelihoods for all pairwise distances. The transitions and transversions parameter makes substitutions of purine by a purine, a pyrimidine by a pyrimidine, a purine by a pyrimidine, or a pyrimidine by a purine equally probable. The rates among sites parameter sets uniform rates by default. Uniform rate assumes variation among sites in the sequences follows a uniform distribution. The pattern among lineages parameter is set by default to homogenous assuming homogeneous substitution patterns among lineages. The data subset parameters are set to partial deletion and 95% which deletes gaps within sequences if there are less than 95% ambiguous sites.

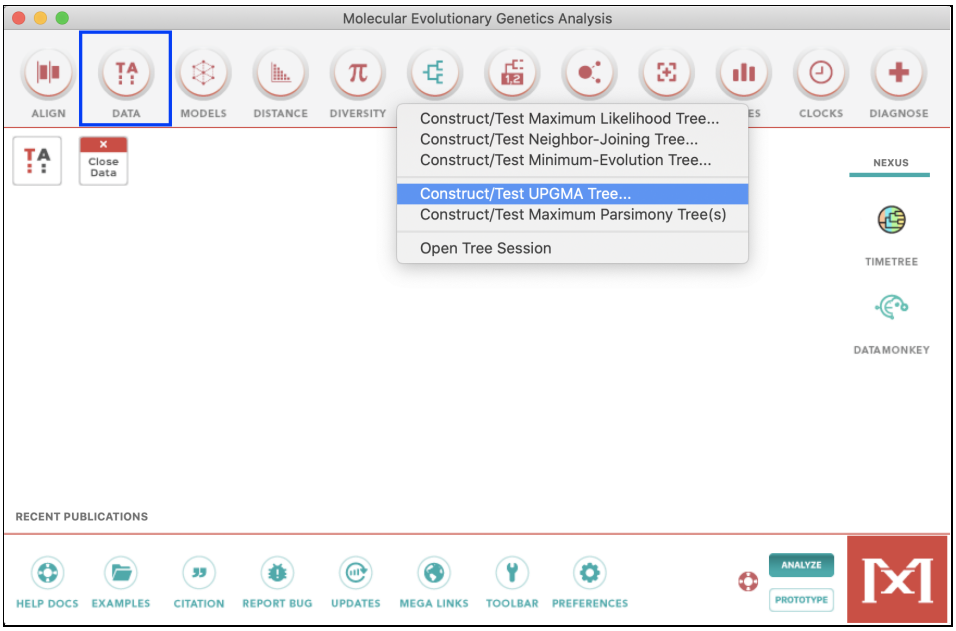


Figure 11. MEGA X GUI interface. First, the "Data" tab is selected to upload the FASTA file for the HYYHCAS motif for human results with spacer length 16. Next, the "Phylogeny" tab is selected to construct a UPGMA tree using select parameters. Adapted from [28].

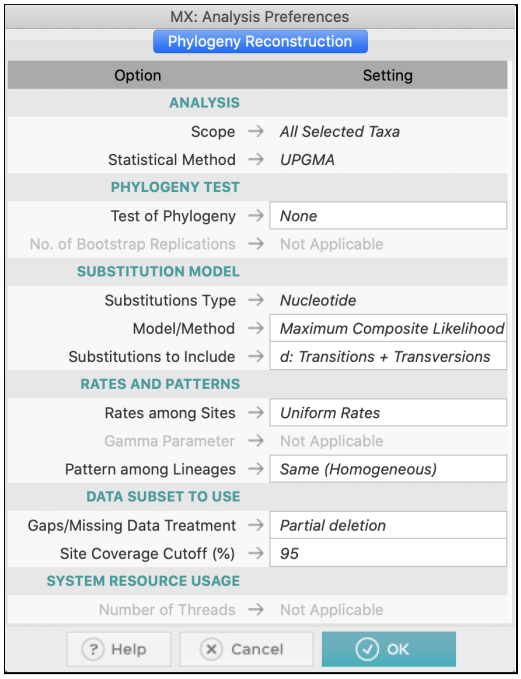


Figure 12. MEGA X parameters used to build the UPGMA tree. Adapted from [28].

4. Subtrees are extracted from the full phylogenetic tree from Step 3 containing the *Hes1* and *Hes5* SPS sequences to check for sequence conservation within the clusters.

5. Build web logos of subtrees using the WebLogo 3 web-based application: http://weblogo.threeplusone.com/create.cgi.

MSA of motifs search TSV results was not performed as the site 1 and site 2 of the SPS is conserved. Performing MSA can disorient the potential SPS sequences such that gaps are inserted preventing proper alignment of site 1 and site 2. To analyze the *Hes1* and *Hes5* SPS sequences, a phylogenetic tree using the MEGA X UPGMA algorithm was used. The UPGMA algorithm is a distance-based method which assumes the rate of evolution of the sequences in the data is the same [29]. UPGMA algorithm can be used for data containing conserved regions such as an SPS. Next, the *Hes1* and *Hes5* SPS sequences were located within the phylogenetic tree, and the subtree containing these and neighboring sequences were extracted and written into a FASTA file to build web logos.

2.2.1 XtractXact Filtered

The *Hes1* and *Hes5* SPS sequences are approximately 25-35 base pairs upstream of a TATA box and since a TATA box is sometimes required for proper activation of Notch target genes, it would be useful to filter the motif search results obtained from the RTGRGAR[N*15-17]HYYHCAS, RTGRGAR[N*15-17]WMMYCAS, and RTGRGAR[N*15-17]YTCHCAY motifs for the human promoter regions to narrow down potential SPSs with a TATA box downstream.

2.3 XtractXact and FIMO Performance Metrics

Users of XtractXact might find performance metrics useful to estimate the time it takes XtractXact to run on their own motifs. Simple performance metric tests were done for

XtractXact and to compare performance with FIMO which outputs similar formatted results. Hmmer was not included for performance metric comparisons as the results were not usable.

XtractXact performance metrics included varying number of spacer lengths (15, 15/16, 15/16/17) on intragenic and promoter regions of all human genes. The promoter regions used included 500, 1500, and 2500 base pairs upstream of all human genes. The motif used for this test was the RTGRGAR[N*15-17]HYYHCAS SPS motif only.

To compare XtractXact and FIMO, various metrics are assessed as shown in Table 8. This includes the time it takes for XtractXact and FIMO to run a varying number of motifs (1, 5, 10, 15) with varying complexity and a spacer length of 16 for both intragenic and promoter regions (500, 1500, 2500 base pairs upstream of all human genes). Three separate FASTA files containing extracted promoter sequences for all hg19 genes were created as input for FIMO, one for each promoter length. This motif search space includes all chromosomes and both strands of the DNA. The varying complexity dimeric motifs were randomly created using the “choices” function within Python’s random module. The choices function returns a list of elements with replacements allowing IUPAC characters to appear more than once randomly. An example of the code is shown in Figure 13. One motif was used as the motif for site 1 of the SPS and subsequent motifs were used for site 2 of the SPS. Table 9 shows the various motifs used for the test. The ggplot2 library in R Studio is used to create bar graphs to visualize test metric comparisons.

Test Metric	**Parameters Tested**			
Number of Motifs	1	5	10	15
Upstream Region	2500bp	1500bp	500bp	

Table 8. Test metrics to compare XtractXact vs. FIMO. Varying number of motifs were tested (1, 5, 10, and 15). Different promoter lengths were tested (2500, 1500, and 500 base pairs upstream of genes).

Motifs Tested	Site 1	N*16	Site 2	Full motif
1	BBYCRGR	NNNNNNNNNNNNNNNN	SSMYGMG	BBYCRGRNNNNNNNNNNNNNNNNSSMYGMG
5	DCYWSYS	NNNNNNNNNNNNNNNN	MNKSGDA	DCYWSYSNNNNNNNNNNNNNNNNMNKSGDA
	DCYWSYS	NNNNNNNNNNNNNNNN	YNARVHG	DCYWSYSNNNNNNNNNNNNNNNNYNARVHG
	DCYWSYS	NNNNNNNNNNNNNNNN	HNBWGTR	DCYWSYSNNNNNNNNNNNNNNNNHNBWGTR
	DCYWSYS	NNNNNNNNNNNNNNNN	MKNBYAY	DCYWSYSNNNNNNNNNNNNNNNNMKNBYAY
	DCYWSYS	NNNNNNNNNNNNNNNN	MNWAGRN	DCYWSYSNNNNNNNNNNNNNNNNMNWAGRN
10	GRVRVWV	NNNNNNNNNNNNNNNN	CHGRTVN	GRVRVWVNNNNNNNNNNNNNNNNCHGRTVN
	GRVRVWV	NNNNNNNNNNNNNNNN	SBGGDAN	GRVRVWVNNNNNNNNNNNNNNNNSBGGDAN
	GRVRVWV	NNNNNNNNNNNNNNNN	VKGRDAD	GRVRVWVNNNNNNNNNNNNNNNNVKGRDAD
	GRVRVWV	NNNNNNNNNNNNNNNN	MCTNKDB	GRVRVWVNNNNNNNNNNNNNNNNMCTNKDB
	GRVRVWV	NNNNNNNNNNNNNNNN	BDRKKKT	GRVRVWVNNNNNNNNNNNNNNNNBDRKKKT
	GRVRVWV	NNNNNNNNNNNNNNNN	RYHACBV	GRVRVWVNNNNNNNNNNNNNNNNRYHACBV
	GRVRVWV	NNNNNNNNNNNNNNNN	BYNWYAG	GRVRVWVNNNNNNNNNNNNNNNNBYNWYAG
	GRVRVWV	NNNNNNNNNNNNNNNN	TTKNRSB	GRVRVWVNNNNNNNNNNNNNNNNTTKNRSB
	GRVRVWV	NNNNNNNNNNNNNNNN	WYMMCGC	GRVRVWVNNNNNNNNNNNNNNNNWYMMCGC
	GRVRVWV	NNNNNNNNNNNNNNNN	GWDKBCS	GRVRVWVNNNNNNNNNNNNNNNNGWDKBCS
15	KBHADYB	NNNNNNNNNNNNNNNN	VBDANDV	KBHADYBNNNNNNNNNNNNNNNNVBDANDV
	KBHADYB	NNNNNNNNNNNNNNNN	KHVKVRA	KBHADYBNNNNNNNNNNNNNNNNKHVKVRA
	KBHADYB	NNNNNNNNNNNNNNNN	BKBMCGC	KBHADYBNNNNNNNNNNNNNNNNBKBMCGC
	KBHADYB	NNNNNNNNNNNNNNNN	WKMHNBV	KBHADYBNNNNNNNNNNNNNNNNWKMHNBV
	KBHADYB	NNNNNNNNNNNNNNNN	CWKWYCV	KBHADYBNNNNNNNNNNNNNNNNCWKWYCV
	KBHADYB	NNNNNNNNNNNNNNNN	SVVMDWM	KBHADYBNNNNNNNNNNNNNNNNSVVMDWM
	KBHADYB	NNNNNNNNNNNNNNNN	SSKBHNB	KBHADYBNNNNNNNNNNNNNNNNSSKBHNB
	KBHADYB	NNNNNNNNNNNNNNNN	DTDBSDY	KBHADYBNNNNNNNNNNNNNNNNDTDBSDY
	KBHADYB	NNNNNNNNNNNNNNNN	TKDTTNW	KBHADYBNNNNNNNNNNNNNNNNTKDTTNW
	KBHADYB	NNNNNNNNNNNNNNNN	KBTWADC	KBHADYBNNNNNNNNNNNNNNNNKBTWADC
	KBHADYB	NNNNNNNNNNNNNNNN	GTDMNAH	KBHADYBNNNNNNNNNNNNNNNNGTDMNAH
	KBHADYB	NNNNNNNNNNNNNNNN	KRWSNWK	KBHADYBNNNNNNNNNNNNNNNNKRWSNWK
	KBHADYB	NNNNNNNNNNNNNNNN	HTCBBHR	KBHADYBNNNNNNNNNNNNNNNNHTCBBHR
	KBHADYB	NNNNNNNNNNNNNNNN	VVRNTRH	KBHADYBNNNNNNNNNNNNNNNNVVRNTRH
	KBHADYB	NNNNNNNNNNNNNNNN	CMWYDYH	KBHADYBNNNNNNNNNNNNNNNNCMWYDYH

Table 9. Motif dataset used to test XtractXact and FIMO. 1, 5, 10, and 15 motifs were used to test various performance metrics.

```
import random
list_iupac = ["A","C","G","T","R","Y","S","W","K","M","B","D","H","V","N"]
print(random.choices(list_iupac,7))
['D', 'C', 'Y', 'W', 'S', 'Y', 'S']
```

Figure 13. Sample Python code used to obtain random motifs used to test XtractXact vs. FIMO.

To summarize, the following methods and performance metrics tests were used.

1. FIMO using the DCYWSYS[N*16]MNKSGDA motif 2500 base pairs upstream of all human genes.
2. Hmmer using variations of the RTGRGAR motif on chromosome 1.
3. XtractXact to extract all three SPS motifs from both intragenic and promoter (2500 base pairs) regions of all human and mouse genes.
4. XtractXact performance metric results.
5. XtractXact vs. FIMO performance metric results.

3. RESULTS

3.1 FIMO Results

FIMO results include the motif_id, sequence_name, start, stop, strand, score, p-value, and matched_sequence columns (Table 10). The motif_id is the name of the file containing the motif searched. The sequence_name column values are the headers for the required FASTA file that FIMO searches for the motif. As shown in Table 10, the partial results show chromosome Y matched motifs. A total of 96,759 motif matches were reported. The start and stop locations indicate the start and stop positions of the matched motif in the chromosome. Strand identifies the DNA strand in which the motif match was found. The score column shows the PWM scoring matrix summed scores for each position of the sequences. The p-value is the probability of a random sequence of the same length as the queried motif sequence matching the position in the genome with an equal or better score calculated by FIMO’s log-odds matrix.

In comparison, XtractXact had found only 939 matches for the same DCYWSYS[N*16]MNKSGDA motif within the 2500 base pair promoter region of all human genes.

motif_id	sequence_name	start	stop	strand	score	p-value	matched_sequence
fimo_motif1_XtractXact_motif2	chrY	28773651	28773680	-	14.416	5.72E-05	GCCTGTCACCTAGGCTGTAGTCCACTGGCA
fimo_motif1_XtractXact_motif2	chrY	28549326	28549355	-	14.832	3.43E-05	GCCTCCCAAAGTGCTGGGACTACAGGCGTA
fimo_motif1_XtractXact_motif2	chrY	27857991	27858020	+	14.832	3.43E-05	GCCTCCCAAAGAGCTGGGATTACAGGCGTA
fimo_motif1_XtractXact_motif2	chrY	27737084	27737113	+	14.416	5.72E-05	ACCACTGGCCGCAGAGGTTTTCAACTGGCA
fimo_motif1_XtractXact_motif2	chrY	27630249	27630278	+	14.416	5.72E-05	CCCTGTGGAGCACTGAGACCAGGCTGGGGA
fimo_motif1_XtractXact_motif2	chrY	27578540	27578569	-	14.416	5.72E-05	CCCAGCCACTCAAGAGGCTGAGGCAGGGGA
fimo_motif1_XtractXact_motif2	chrY	27537442	27537471	-	14.416	5.72E-05	TCTTGTCACCCAGGCTGGAGTGCAGTGGCA
fimo_motif1_XtractXact_motif2	chrY	27531408	27531437	+	14.416	5.72E-05	TCTTGTCACCCAGGCTGGAGTGCAGTGGCA
fimo_motif1_XtractXact_motif2	chrY	27165577	27165606	+	14.832	3.43E-05	TCCTCTCATTTGATTAACACTTGCATGGAA
fimo_motif1_XtractXact_motif2	chrY	26425439	26425468	-	14.416	5.72E-05	TCTTGTCACCCAGGCTGGAGTGCAGTGGCA
fimo_motif1_XtractXact_motif2	chrY	26380219	26380248	-	14.416	5.72E-05	CCCAGCCACTCAAGAGGCTGAGGCAGGGGA
fimo_motif1_XtractXact_motif2	chrY	26330773	26330802	+	14.416	5.72E-05	CCCTGTGGAGCACTGAGACCAGGCTGGGGA
fimo_motif1_XtractXact_motif2	chrY	26224551	26224580	+	14.416	5.72E-05	ACCACTGGCCGCAGAGGTTTTCAACTGGCA
fimo_motif1_XtractXact_motif2	chrY	26104649	26104678	+	14.832	3.43E-05	GCCTCCCAAAGAGCTGGGATTACAGGCGTA
fimo_motif1_XtractXact_motif2	chrY	24479776	24479805	-	14.832	3.43E-05	GCCACCGTACCCAGCAGTGGGATAATGGAA
fimo_motif1_XtractXact_motif2	chrY	23630935	23630964	-	14.416	5.72E-05	CCTTCCCGCACAGCGGAAAGTTAAAGGGTA
fimo_motif1_XtractXact_motif2	chrY	22918374	22918403	-	14.832	3.43E-05	TCCTCCCTCCCTCTGTGTCTGTACAGGGGA
fimo_motif1_XtractXact_motif2	chrY	21181227	21181256	+	14.416	5.72E-05	TCCAGCCCAGTGGTGTCCTCACCCAGGGCA
fimo_motif1_XtractXact_motif2	chrY	21153430	21153459	-	14.832	3.43E-05	GCCACTGTGGAGGTGGGGCAGTTACTGGGA
fimo_motif1_XtractXact_motif2	chrY	21095255	21095284	-	14.832	3.43E-05	TCCTGCGTAAGGGAGCACACGACCCTGGAA
fimo_motif1_XtractXact_motif2	chrY	21095688	21095717	+	14.416	5.72E-05	CCTTCCCGCTGCGGTAGCGCCGCCGGGGAA
fimo_motif1_XtractXact_motif2	chrY	21096095	21096124	-	14.416	5.72E-05	CCCAGTCGGTCCCCCACAGCCCCCTGGGGA
fimo_motif1_XtractXact_motif2	chrY	21096799	21096828	-	14.832	3.43E-05	TCCTGCGTAAGGGAGCACACGACCCTGGAA
fimo_motif1_XtractXact_motif2	chrY	21094837	21094866	-	14.832	3.43E-05	TCCTGCGTAAGGGAGCACACGACCCTGGAA
fimo_motif1_XtractXact_motif2	chrY	21095270	21095299	+	14.416	5.72E-05	CCTTCCCGCTGCGGTAGCGCCGCCGGGGAA
fimo_motif1_XtractXact_motif2	chrY	21095677	21095706	-	14.416	5.72E-05	CCCAGTCGGTCCCCCACAGCCCCCTGGGGA
fimo_motif1_XtractXact_motif2	chrY	21035305	21035334	-	14.832	3.43E-05	TCCTGCGTAAGGGAGCACACGACCCTGGAA
fimo_motif1_XtractXact_motif2	chrY	21035738	21035767	+	14.416	5.72E-05	CCTTCCCGCTGCGGTAGCGCCGCCGGGGAA
fimo_motif1_XtractXact_motif2	chrY	21036145	21036174	-	14.416	5.72E-05	CCCAGTCGGTCCCCCACAGCCCCCTGGGGA
fimo_motif1_XtractXact_motif2	chrY	21035331	21035360	-	14.832	3.43E-05	TCCTGCGTAAGGGAGCACACGACCCTGGAA
fimo_motif1_XtractXact_motif2	chrY	21035764	21035793	+	14.416	5.72E-05	CCTTCCCGCTGCGGTAGCGCCGCCGGGGAA
fimo_motif1_XtractXact_motif2	chrY	21036171	21036200	-	14.416	5.72E-05	CCCAGTCGGTCCCCCACAGCCCCCTGGGGA
fimo_motif1_XtractXact_motif2	chrY	21036849	21036878	-	14.832	3.43E-05	TCCTGCGTAAGGGAGCACACGACCCTGGAA
fimo_motif1_XtractXact_motif2	chrY	20995776	20995805	+	14.416	5.72E-05	CCTTCTGAAAGAGGACACTTATGCTGGGAA

Table 10. Partial FIMO results for the DCYWSYS[N*16]MNKSGDA motif.

3.2 HMMER Results

Results are shown in Figure 14 after running the nhmmer command. Results do not contain any potential motif hits against chromosome 1 using the site 1 RTGRGAR motif variations, though many residues had passed the various filters. The RTGRGAR motif exists in chromosome 1, but hmmer did not report any of the matches. This indicates a potential problem within the underlying hmmer algorithm for this specific project goal of extracting SPS motifs from the human genome.

```
# nhmmer :: search a DNA model, alignment, or sequence against a DNA database
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query file:                   first_site.hmm
# target sequence database:     chr1.fa
# sequence reporting threshold:   E-value <= 10
# Max sensitivity mode:         on [all heuristic filters off]
# input query is asserted as:   DNA
# window length :               7
# number of worker threads:     2
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       motif1  [M=7]
Scores for complete hits:
   E-value  score  bias  Sequence  start    end  Description
   ------- ------ -----  --------  -----  -----  -----------

  [No hits detected that satisfy reporting thresholds]

Annotation for each hit  (and alignments):

  [No targets detected that satisfy reporting thresholds]

Internal pipeline statistics summary:
-------------------------------------
Query model(s):                         1  (7 nodes)
Target sequences:                       1  (498501242 residues searched)
Residues passing SSV filter:        20676999  (0.0415); expected (0.02)
Residues passing bias filter:       20676999  (0.0415); expected (0.02)
Residues passing Vit filter:        20676999  (0.0415); expected (1)
Residues passing Fwd filter:        20676999  (0.0415); expected (1)
Total number of hits:                   0  (0)
# CPU time: 220.09u 0.21s 00:03:40.30 Elapsed: 00:01:49.81
# Mc/sec: 31.78
//
[ok]
```

Figure 14. Results against chromosome 1. No hits detected.

3.3 XtractXact Results

XtractXact results files are listed under the "XtractXact Motif Results Filenames" column in Table 11. Each of these files contain multiple columns and partial results of the human promoter for the RTGRGAR[N*15-17]HYYHCAS motif is shown in Table 12. The Motif2TATA column produced by XtractXact indicates the nucleotide length from the end of the extracted SPS motif to a potential TATA box downstream. If a TATA box does not exist downstream, the given value is False. The TATA2Gene column indicates the nucleotide length from the beginning of the TATA box to the start site of the downstream gene. If a TATA box does not exist, this column value is also False. The Motif2Gene column indicates the nucleotide length between the end of the potential SPS motif to the transcription start site of the gene.

Both intragenic and promoter results for each tested motif for human and mouse were split into various files separated by spacer length (Table 11). The exact *Hes1* and *Hes5* SPS sequences with exact start and stop locations were found in the XtractXact results as shown in Table 13. The UCSC genome coordinates for the SPS and the distances between the motif and TATA box were thus validated as shown in Figure 15 and 16.

Species	XtractXact Motif Results Filenames	Spacer Lengths		
		15	16	17
Human	XtractXact_RTGRGAR_HYYHCAS_human_intragenic.txt	HYYHCAS_intragenic_human_15.tsv	HYYHCAS_intragenic_human_16.tsv	HYYHCAS_intragenic_human_17.tsv
	XtractXact_RTGRGAR_HYYHCAS_human_promoter.txt	HYYHCAS_promoter_human_15.tsv	HYYHCAS_promoter_human_16.tsv	HYYHCAS_promoter_human_17.tsv
	XtractXact_RTGRGAR_WYYMCAS_human_intragenic.txt	WYYMCAS_intragenic_human_15.tsv	WYYMCAS_intragenic_human_16.tsv	WYYMCAS_intragenic_human_17.tsv
	XtractXact_RTGRGAR_WYYMCAS_human_promoter.txt	WYYMCAS_promoter_human_15.tsv	WYYMCAS_promoter_human_16.tsv	WYYMCAS_promoter_human_17.tsv
	XtractXact_RTGRGAR_YTCHCAY_human_intragenic.txt	YTCHCAY_intragenic_human_15.tsv	YTCHCAY_intragenic_human_16.tsv	YTCHCAY_intragenic_human_17.tsv
	XtractXact_RTGRGAR_YTCHCAY_human_promoter.txt	YTCHCAY_promoter_human_15.tsv	YTCHCAY_promoter_human_16.tsv	YTCHCAY_promoter_human_17.tsv
Mouse	XtractXact_RTGRGAR_HYYHCAS_mouse_intragenic.txt	HYYHCAS_intragenic_mouse_15.tsv	HYYHCAS_intragenic_mouse_16.tsv	HYYHCAS_intragenic_mouse_17.tsv
	XtractXact_RTGRGAR_HYYHCAS_mouse_promoter.txt	HYYHCAS_promoter_mouse_15.tsv	HYYHCAS_promoter_mouse_16.tsv	HYYHCAS_promoter_mouse_17.tsv
	XtractXact_RTGRGAR_WYYMCAS_mouse_intragenic.txt	WYYMCAS_intragenic_mouse_15.tsv	WYYMCAS_intragenic_mouse_16.tsv	WYYMCAS_intragenic_mouse_17.tsv
	XtractXact_RTGRGAR_WYYMCAS_mouse_promoter.txt	WYYMCAS_promoter_mouse_15.tsv	WYYMCAS_promoter_mouse_16.tsv	WYYMCAS_promoter_mouse_17.tsv
	XtractXact_RTGRGAR_YTCHCAY_mouse_intragenic.txt	YTCHCAY_intragenic_mouse_15.tsv	YTCHCAY_intragenic_mouse_16.tsv	YTCHCAY_intragenic_mouse_17.tsv
	XtractXact_RTGRGAR_YTCHCAY_mouse_promoter.txt	YTCHCAY_promoter_mouse_15.tsv	YTCHCAY_promoter_mouse_16.tsv	YTCHCAY_promoter_mouse_17.tsv

Table 11. All files created for each site 2 motif and spacer lengths.

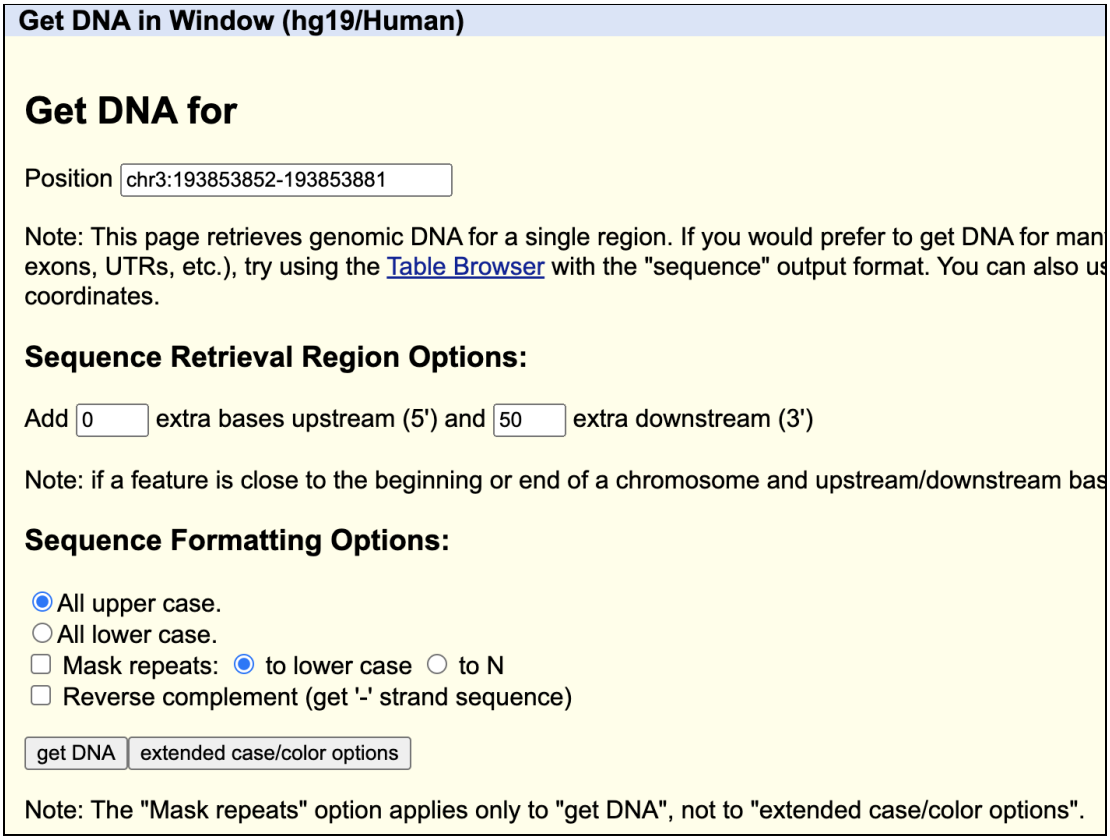
Chromosome	Strand	Gene	Gene Start	Gene Stop	Accession ID	Motif1	Spacer Length	Motif2	Motif Match	Match Start	Match Stop	Motif2TATA	TATA2Gene	Motif2Gene
chr17	-	AA06	31856806	31860779	NR_037584.2	RTGRGAR	16	HYYHCAS	ATGGGAACAATCCTAATACTTACCCCACAG	31862021	31862050	False	False	1242
chr2	-	ABCA12	215796266	215896810	NM_015657.4	RTGRGAR	16	HYYHCAS	GTGGGAAGCCATATGCTACATTTTTCCAC	215899160	215899189	261	2089	2350
chr7	+	ABCB8	150725536	150744869	NM_001282393.2	RTGRGAR	17	HYYHCAS	GTGGGAAGGAAAGACCTGACCGTCCCCCAG	150724489	150724519	False	False	1018
chr3	+	ABCF3	183903862	183911795	NM_001351298.1	RTGRGAR	15	HYYHCAS	GTGGGAGGGTTGGTTGCCCCTCACCTCAG	183901724	183901752	1383	726	2111
chr3	+	ABCF3	183903986	183911793	NM_001351299.2	RTGRGAR	15	HYYHCAS	GTGGGAGGGTTGGTTGCCCCTCACCTCAG	183901724	183901752	1383	850	2235
chr2	+	ABHD1	27346682	27353680	NM_032604.4	RTGRGAR	16	HYYHCAS	ATGAGAATCTAATTCCACTGCTGATCTCAG	27344574	27344603	259	1819	2080
chr6	+	ABRACL	139349881	139364439	NM_021243.3	RTGRGAR	16	HYYHCAS	ATGAGAACTGTTTGCTCCAGGTCACCTCAC	139349063	139349092	False	False	790
chr15	+	ACAN	89346666	89418584	NM_0011354	RTGRGAR	17	HYYHCAS	ATGAGAATCCATGACCTCGTTGGAATTTCAG	89344712	89344742	False	False	1925
chr3	-	ACAP2	194995474	195163749	NM_0122876	RTGRGAR	17	HYYHCAS	ATGAGAAGATGTACATGGGAGCAATTTCCAG	195164355	195164385	False	False	606
chr9	+	ACO1	32384640	32454767	NM_0021973	RTGRGAR	17	HYYHCAS	GTGAGAAGGGAGGCGGGCCAAGTGACCACAC	32384231	32384261	False	False	380
chr20	-	ACOT8	44470360	44486031	NM_0054694	RTGRGAR	15	HYYHCAS	ATGGGAGAAAACTGCAGTTATAAATTCCAG	44486845	44486873	False	False	814
chrX	-	ACOT9	23719172	23761393	NM_0010371712	RTGRGAR	15	HYYHCAS	GTGGGAAGCTCCGCAGAGGCGGACTCCAG	23761476	23761504	False	False	83
chr4	-	ACSL1	185676749	185747122	NM_0013818781	RTGRGAR	15	HYYHCAS	GTGGGAGACACCCAGGGAAAGTCCTTCAG	185748335	185748363	False	False	1213
chr6	+	ACTG1P9	46172648	46174298	ACTG1P9	RTGRGAR	16	HYYHCAS	GTGAGAAGTCCAAGATCAAGGTGCCTTCAG	46171703	46171732	627	288	917
chr12	+	ACVRL1	52301287	52317145	NM_000020.3	RTGRGAR	15	HYYHCAS	ATGAGAATCTGCTTGAGGTTTCACCCCAC	52298795	52298823	False	False	2465
chr12	+	ACVRL1	52301287	52317145	NM_000020.3	RTGRGAR	17	HYYHCAS	ATGAGAATCTGCTTGAGGTTTCACCCCAC	52298795	52298825	False	False	2463
chr18	+	ADAD1P2	47590161	47590856	ADAD1P2	RTGRGAR	15	HYYHCAS	ATGGGAACAGAGAAGCTCAGCGCTTCCAG	47589524	47589552	False	False	610
chr15	+	ADAL	43622548	43641613	NM_0013243642	RTGRGAR	16	HYYHCAS	ATGGGAGGCTGGATGTCTACTGACTCCCAG	43620940	43620969	False	False	1580
chr15	+	ADAL	43622869	43641613	NM_0013243652	RTGRGAR	16	HYYHCAS	ATGGGAGGCTGGATGTCTACTGACTCCCAG	43620940	43620969	False	False	1901
chr15	+	ADAL	43622869	43646767	NM_0011592803	RTGRGAR	16	HYYHCAS	ATGGGAGGCTGGATGTCTACTGACTCCCAG	43620940	43620969	False	False	1901
chr2	-	ADAM17	9628615	9695959	NM_0013827781	RTGRGAR	16	HYYHCAS	GTGGGAAGATTCCAGCTCTTGTACTCACAG	9696720	9696749	False	False	761
chr7	+	ADCY1	45613738	45704279	NM_0012817682	RTGRGAR	15	HYYHCAS	ATGAGAATGACATGTAAACCACACCACAC	45611434	45611462	False	False	2277
chr2	-	ADD2	70883916	70994847	NM_0011850542_2	RTGRGAR	15	HYYHCAS	GTGAGAGTTGGAGAGCGTTGGGTCCACAG	70995888	70995916	False	False	1041
chr2	-	ADD2	70883916	70995332	NM_017488.4_2	RTGRGAR	15	HYYHCAS	GTGAGAGTTGGAGAGCGTTGGGTCCACAG	70995888	70995916	False	False	556

Table 12. Partial results from XtractXact for RTGRGAR[N*15,16,17]HYYHCAS motif for human promoter.

Chromosome	Strand	Gene	…	Motif Match	Match Start	Match Stop	Motif2TATA	TATA2Gene	Motif2Gene
chr10	+	HEAT2	…	GTGAGAGTCAGTCTCAGGCCAGATTCCAC	112871825	112871853	False	False	2196
chr3	**+**	**HES1**	**…**	**GTGGGAAAGAAAGTTTGGGAAGTTTCACAC**	**193853852**	**193853881**	**24**	**30**	**56**
chr1	-	HES4	…	GTGGGAAAGAATGCGGAGCCGGGTTCACAC	935525	935554	False	False	48
chr1	**-**	**HES5**	**…**	**GTGGGAACGGCCGCGGCGCCCGGACTCCAG**	**2461754**	**2461783**	**23**	**29**	**52**
chr4	+	HGFAC	…	GTGGGAGGCAGCCAGGGTTAATCATTTCCAC	3443626	3443656	False	False	39

Table 13. *Hes1* and *Hes5* found in XtractXact results with the correct UCSC genome coordinates.

A.

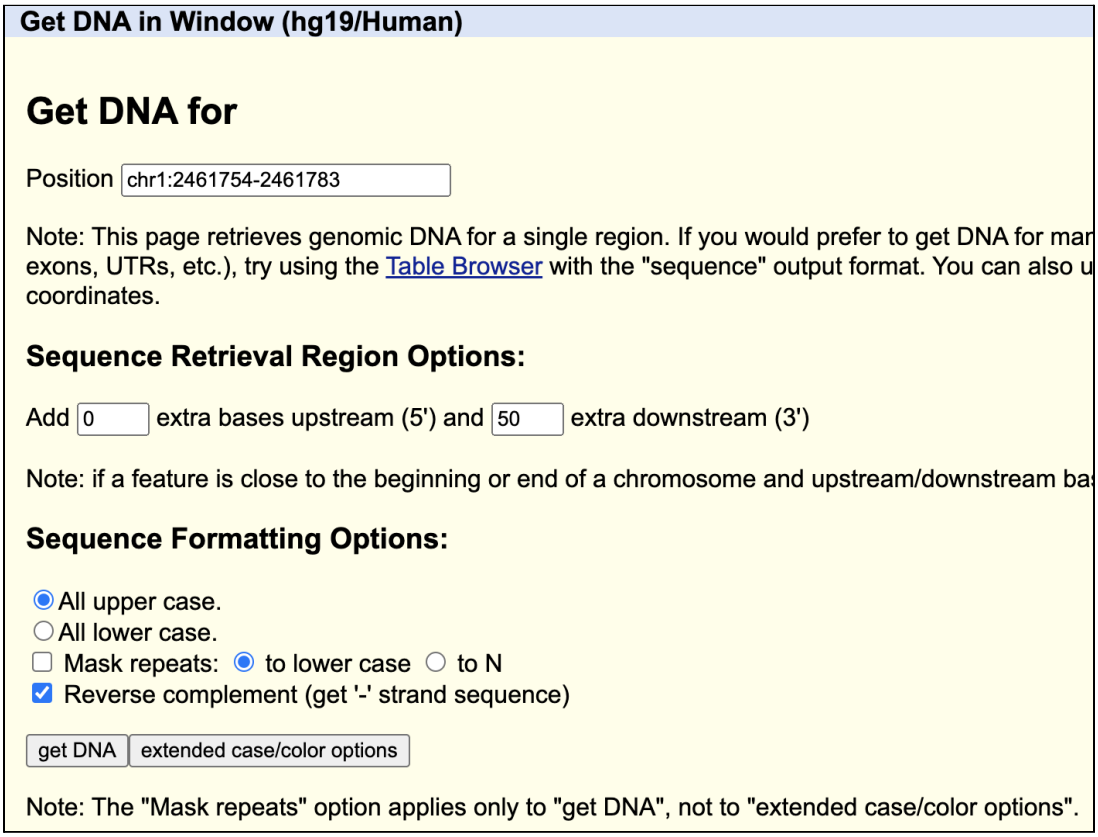


B.

```
>hg19_dna range=chr3:193853852-193853931 5'pad=0 3'pad=50 strand=+ repeatMasking=none
GTGGGAAAGAAAGTTTGGGAAGTTTCACACGAGCCGTTCGCGTGCAGTCC
CAGATATATATAGAGGCCGCCAGGGCCTAG
```

Figure 15. A. Genomic coordinates for *Hes1* including 50 base pairs downstream to verify correct XtractXact results. B. Genomic sequence returned using *Hes1* coordinates. The TATA box location is 24 base pairs from the SPS motif stop location.

A.



B.

```
>hg19_dna range=chr1:2461704-2461783 5'pad=0 3'pad=50 strand=- repeatMasking=none
GTGGGAACGGCCGCGGCGCCCGGACTCCAGGCGCCAGGCCGCCGCCCGCG
CCTATATAGGGCGTCGGCGCGCGGGGCCGG
```

Figure 16. A. Genomic coordinates for *Hes5* including 50 base pairs downstream to verify correct XtractXact results. Reverse complement is checked to get the negative strand. B. Genomic sequence returned using *Hes5* coordinates. The TATA box location is 23 base pairs from the SPS motif stop location.

Similarly, XtractXact was able to identify thousands of potential SPS motif matches for the promoter and intragenic regions (Table 14) in the mouse genome of which partial results are shown in Table 15. Table 15 results include identical SPSs for the same gene. This is due to some genes having multiple transcripts that vary in genomic location. The initial XtractXact files were modified to include only one start and stop position per gene for which number of matches are shown in Table 14. The entire genomic sequence (including introns and exons) are included in the motif search.

Species	SPS Site 1 Motif	Spacer Length	SPS Site 2 Motif	Promoter	Intragenic
Human	RTGRGAR	15,16,17	HYYHCAS	1,589	20,416
	RTGRGAR	15,16,17	WYYMCAS	663	8,138
	RTGRGAR	15,16,17	YTCHCAY	251	3,696
Mouse	RTGRGAR	15,16,17	HYYHCAS	1,402	19,219
	RTGRGAR	15,16,17	WYYMCAS	590	8,298
	RTGRGAR	15,16,17	YTCHCAY	218	2,967

Table 14. Number of SPS motif matches found within the promoter and intragenic regions for each site 2 motifs for spacer lengths 15, 16, and 17 combined. Results filtered to include only unique SPSs per gene.

Web logos were created for each of the split spacer length files. The spacer region for SPSs are not known to be conserved and web logos can be used to further validate this claim. As expected, there were no conserved nucleotides within the spacer region for all three SPS motif results as shown in Table 16 and 17 for the human and mouse promoter and intragenic regions. All motif matches from the RTGRGAR[N*16]HYYHCAS SPS motif only were clustered. Figures 17 and 18 show the subtree clusters of the RTGRGAR[N*16]HYYHCAS motif containing *Hes1* and *Hes5* SPS motifs.. Clustering sequences sharing highly similar nucleotides can be used to identify any important functional features or motifs not easily seen by sequence gazing alone. This can reveal which exact motifs are clustered together that cannot be done using a web logo. SPS motifs for other human genes clustered with *Hes1* and *Hes5* can be further processed downstream to identify any distinguishable features such as common gene pathways they are involved in.

Chromosome	Strand	Gene	Gene Start	Gene Stop	Accession ID	Motif1	Spacer Length	Motif2	Motif Match	Match Start	Match Stop	Motif2TATA	TATA2Gene	Motif2Gene
chr7	+	ABCA14	120206096	120325354	XM_006508143.3	RTGRGAR	17	HYYHCAS	ATGAGAAAGTAAATTCTTTATTAATCTCACAG	120205076	120205106	False	False	991
chr10	+	ABCA7	79997614	80015572	NM_013850.1	RTGRGAR	16	HYYHCAS	GTGGGAAAAATGGAACCACCGATCCTCACAG	79996733	79996762	False	False	853
chr10	+	ABCA7	79998450	80015572	XM_017313958.1	RTGRGAR	16	HYYHCAS	GTGGGAAAAATGGAACCACCGATCCTCACAG	79996733	79996762	False	False	1689
chr11	-	ABCA8A	110025634	110072217	XM_017314485.1	RTGRGAR	15	HYYHCAS	ATGGTGAGAAAAAGGATTTTCCCTCTCCAG	110073309	110073337	282	810	1092
chr1	-	ABCB6	75171640	75180392	NM_023732.3	RTGRGAR	16	HYYHCAS	GTGAGAGCCCTAGCTCCCAGCTCCCCTCAC	75181306	75181335	176	738	914
chr1	-	ABCB6	75171715	75180271	XM_006496552.2	RTGRGAR	16	HYYHCAS	GTGAGAGCCCTAGCTCCCAGCTCCCCTCAC	75181306	75181335	176	859	1035
chr2	-	ABHD12	150832493	150904665	XM_006500367.3	RTGRGAR	16	HYYHCAS	ATGGGAGGCAGAGGCAGGCGGATTTCTCAG	150905330	150905359	100	565	665
chr2	-	ABHD12	150832515	150904731	NM_024465.3	RTGRGAR	16	HYYHCAS	ATGGGAGGCAGAGGCAGGCGGATTTCTCAG	150905330	150905359	100	499	599
chr2	+	ABHD16B	181493205	181494980	NM_183181.2	RTGRGAR	16	HYYHCAS	ATGAGAAAATACACCTTCTGACTCTTACAG	181492705	181492734	False	False	472
chr19	+	ABHD17B	21653226	21685638	XM_011247228.2	RTGRGAR	17	HYYHCAS	GTGGGAGGAAATTATACTTTACTATTCCCAC	21651334	21651364	False	False	1863
chr19	+	ABHD17B	21653308	21685637	NM_146096.3	RTGRGAR	17	HYYHCAS	GTGGGAGGAAATTATACTTTACTATTCCCAC	21651334	21651364	False	False	1945
chr5	+	ABLIM2	35814268	35884973	XM_006503925.3	RTGRGAR	17	HYYHCAS	ATGGGAAGTGGGTGGTTGTCATATATCTCAC	35811946	35811976	False	False	2293
chr5	+	ABLIM2	35814274	35884973	XM_006503926.3	RTGRGAR	17	HYYHCAS	ATGGGAAGTGGGTGGTTGTCATATATCTCAC	35811946	35811976	False	False	2299
chr5	+	ABLIM2	35814285	35884973	XM_017320836.1	RTGRGAR	17	HYYHCAS	ATGGGAAGTGGGTGGTTGTCATATATCTCAC	35811946	35811976	False	False	2310
chr5	+	ABLIM2	35814311	35884973	XM_017320837.1	RTGRGAR	17	HYYHCAS	ATGGGAAGTGGGTGGTTGTCATATATCTCAC	35811946	35811976	False	False	2336
chrX	+	ACOT9	155262442	155297654	NM_019736.4	RTGRGAR	15	HYYHCAS	GTGGGAAATTCCATTGTCTCGGTCTCCAG	155262337	155262365	False	False	78
chr11	-	ACOX1	116171883	116199045	NM_001271898.1	RTGRGAR	16	HYYHCAS	GTGGGAGGACTGAAGGAATCTCCACCCCAC	116201064	116201093	1206	813	2019
chr11	-	ACOX1	116172650	116198861	XM_006532002.3	RTGRGAR	16	HYYHCAS	GTGGGAGGACTGAAGGAATCTCCACCCCAC	116201064	116201093	1206	997	2203
chr5	+	ACOX3	35582963	35611887	XM_006504196.3	RTGRGAR	16	HYYHCAS	GTGGGAGGCCCTGTGACAACTGCACCCCAG	35581952	35581981	False	False	983
chr5	+	ACOX3	35582965	35612107	XM_006504200.3	RTGRGAR	16	HYYHCAS	GTGGGAGGCCCTGTGACAACTGCACCCCAG	35581952	35581981	False	False	985
chr5	+	ACOX3	35582965	35613801	XM_017321181.1	RTGRGAR	16	HYYHCAS	GTGGGAGGCCCTGTGACAACTGCACCCCAG	35581952	35581981	False	False	985
chr5	+	ACOX3	35582965	35614208	XM_006504199.3	RTGRGAR	16	HYYHCAS	GTGGGAGGCCCTGTGACAACTGCACCCCAG	35581952	35581981	False	False	985
chr5	+	ACOX3	35583059	35613801	NM_030721.2	RTGRGAR	16	HYYHCAS	GTGGGAGGCCCTGTGACAACTGCACCCCAG	35581952	35581981	False	False	1079
chr5	-	ACTB	142903115	142906754	NM_007393.5	RTGRGAR	17	HYYHCAS	GTGGGAAAGATAACTAGGGGTGTGTTTCCAG	142908151	142908181	False	False	1397

Table 15. Partial results from XtractXact for RTGRGAR[N*15,16,17]HYYHCAS motif for mouse promoter.

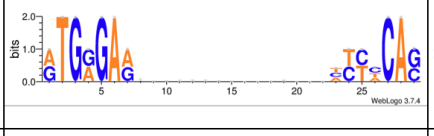
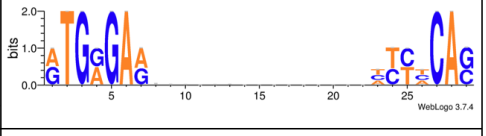
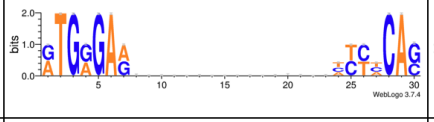
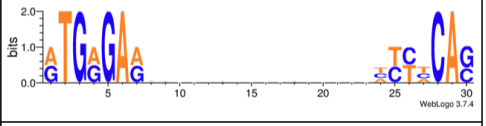
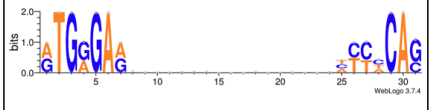
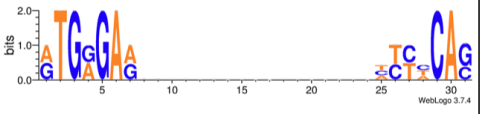
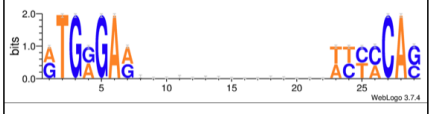
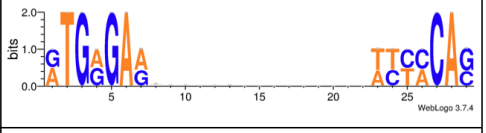
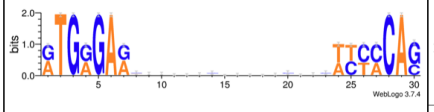
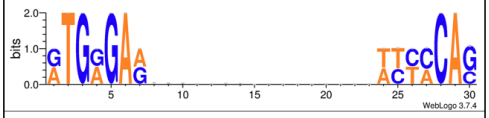
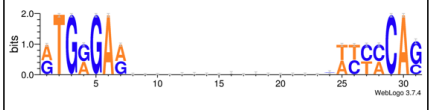
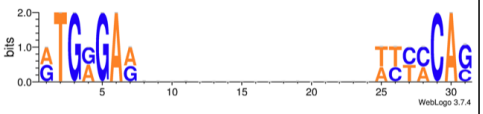
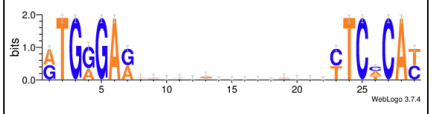
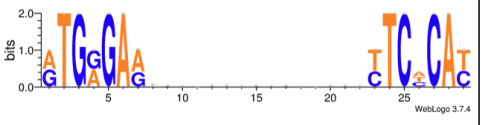
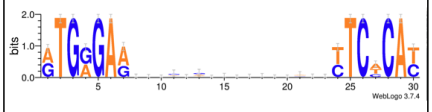
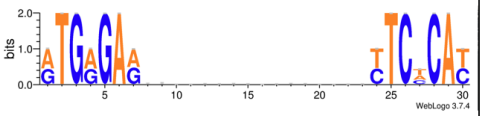
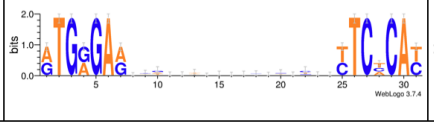
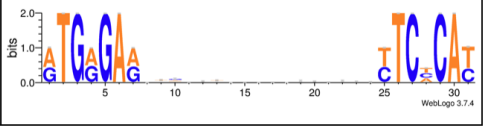
Motif2	Spacer Length	Promoter (2500bp)	Intragenic
HYYHCAS		**2,078 total sequences**	**34,443 total sequences**
	15		
	16		
	17		
WYYMCAS		**873 total sequences**	**13,872 total sequences**
	15		
	16		
	17		
YTCHCAY		**305 total sequences**	**6,226 total sequences**
	15		
	16		
	17		

Table 16. Logos across various motifs, spacer lengths, promoter, and intragenic regions for human. Individual logos created using Web Logo 3.

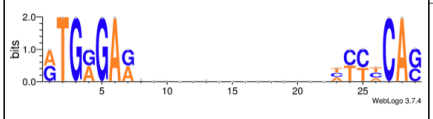
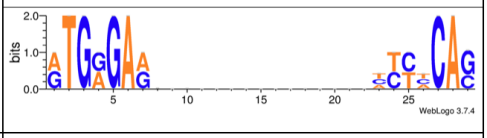
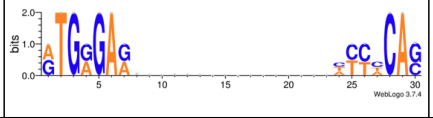
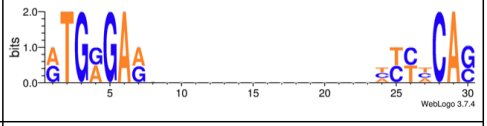
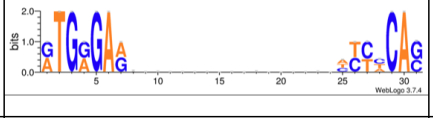
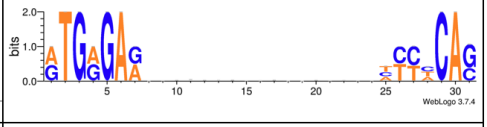
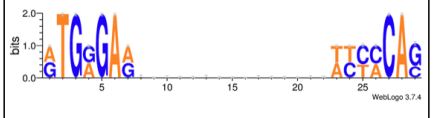
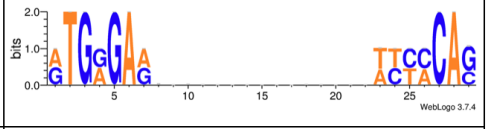
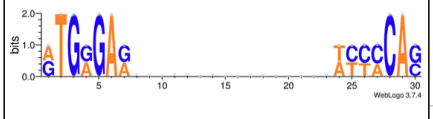
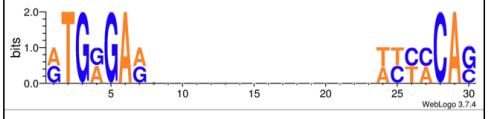
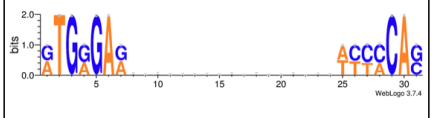
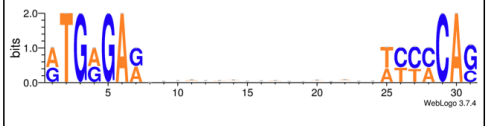
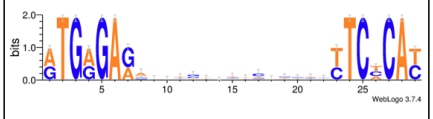
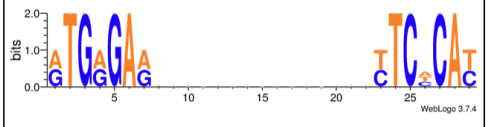
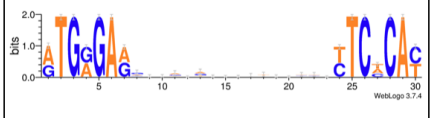
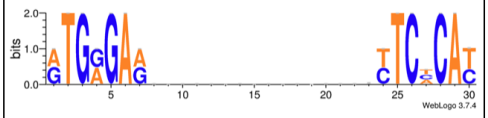
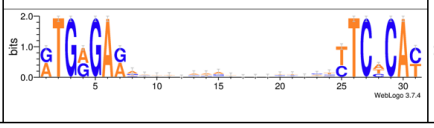
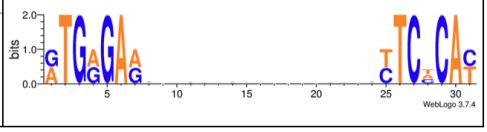
Motif2	Spacer Length	Promoter (2500bp)	Intragenic
HYYHCAS		**3,336 total sequences**	**65,223 total sequences**
	15		
	16		
	17		
WYYMCAS		**1,414 total sequences**	**27,981 total sequences**
	15		
	16		
	17		
YTCHCAY		**494 total sequences**	**10,383 total sequences**
	15		
	16		
	17		

Table 17. Logos across various motifs, spacer lengths, promoter, and intragenic regions for mouse. Individual logos created using Web Logo 3.

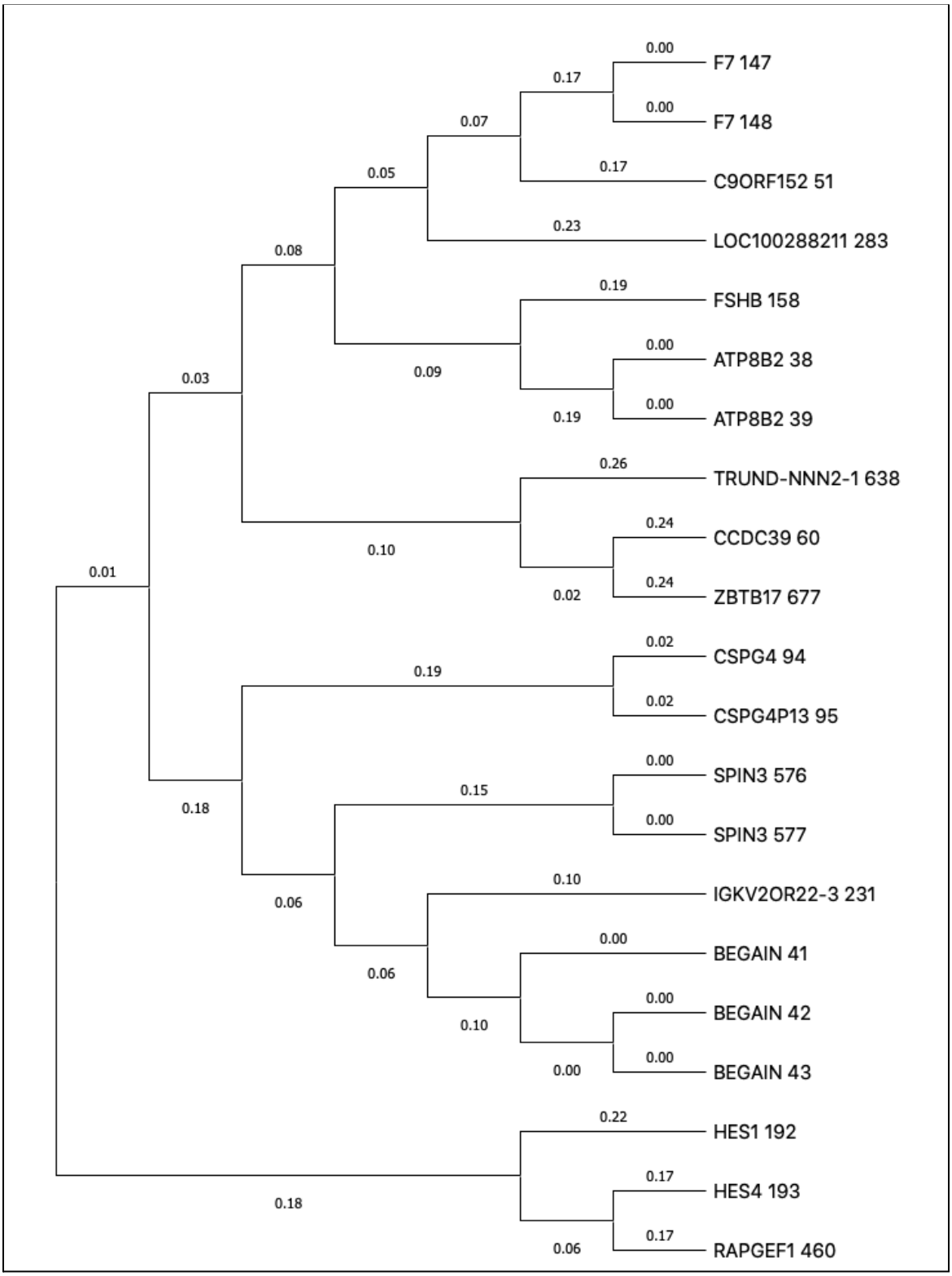


Figure 17. One subtree cluster of the full RTGRGAR[N*16]HYYHCAS containing the *Hes1* SPS motif. The numbers on each branch indicate branch length.

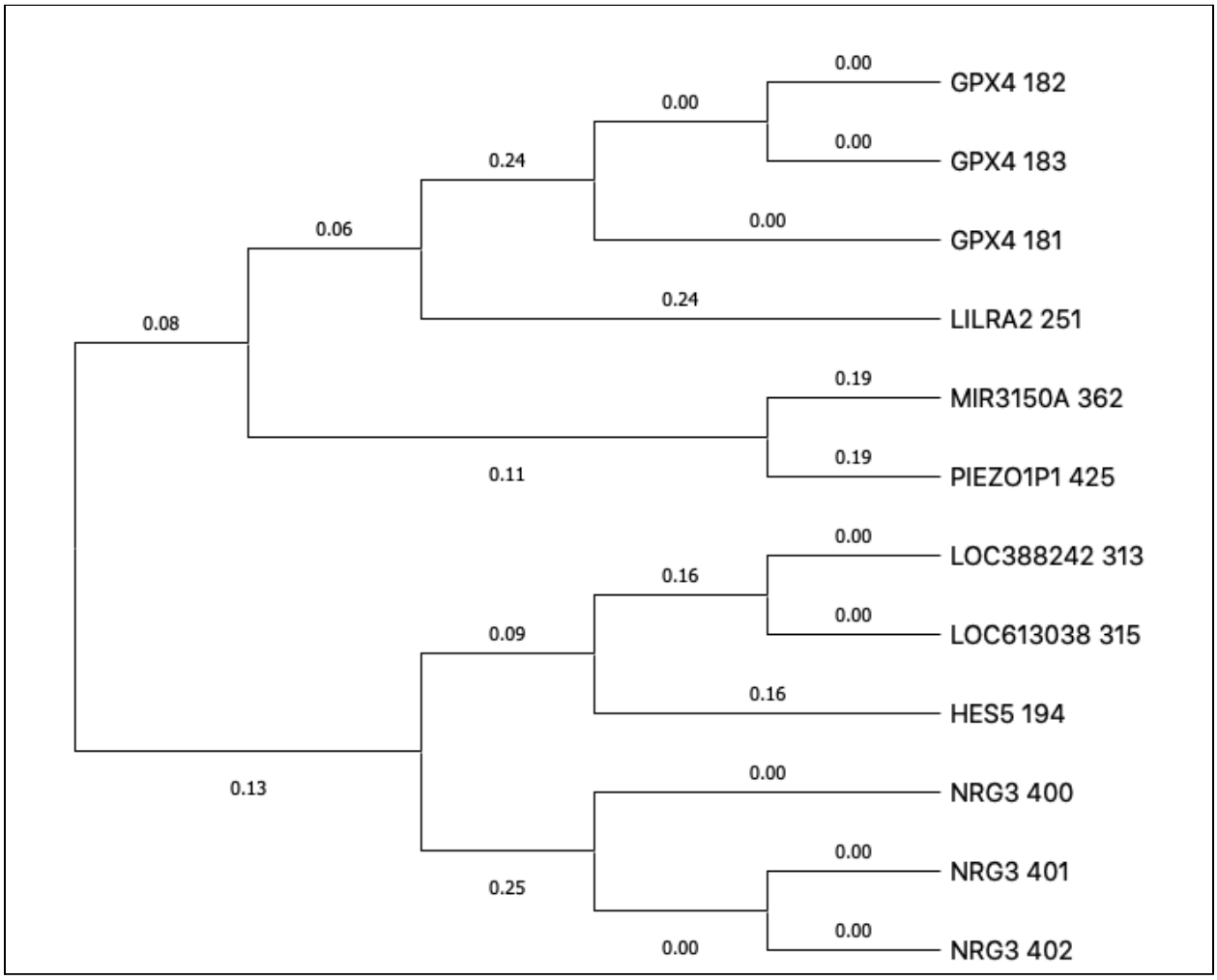


Figure 18. One subtree cluster of the full RTGRGAR[N*16]HYYHCAS containing the *Hes5* SPS motif. The numbers on each branch indicate branch length.

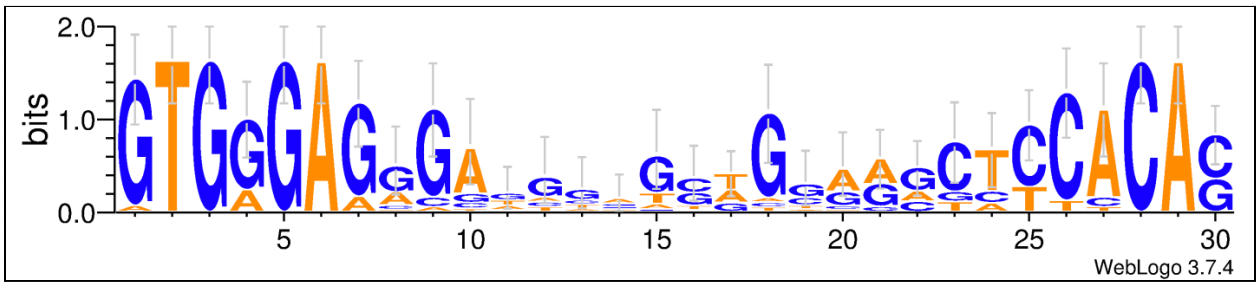


Figure 19. Logo for the subtree cluster (Figure 17) containing the *Hes1* SPS motif (GTGGGAA[N*16]TTCACAC. Logo created using Web Logo 3.

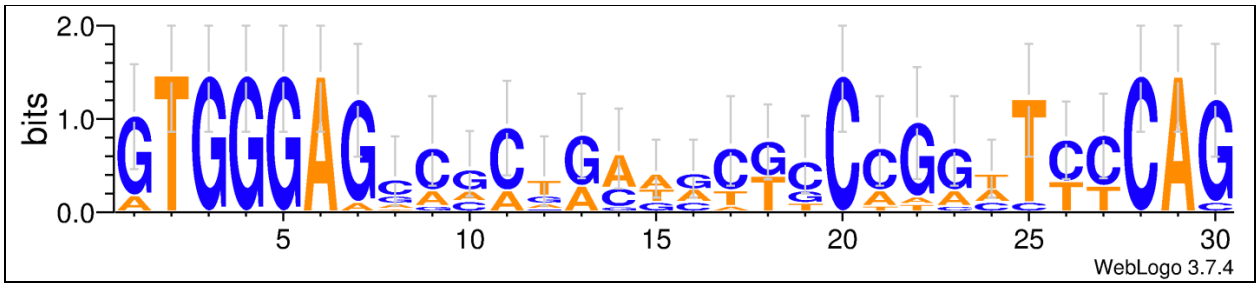


Figure 20. Logo for the subtree cluster (Figure 18) containing the *Hes5* SPS motif (GTGGGAA[N*16]ACTCCAG). Logo created using Web Logo 3.

Figures 19 and 20 show the subtree cluster web logos containing the *Hes1* and *Hes5* SPS. The *Hes1* subtree web logo shows conserved nucleotides GTGGGA for site 1 and CACA for site 2. For the *Hes5* subtree web logo, the conserved nucleotides were GTGGGAG for site 1 and CAG for site 2. Site 1 for both web logos show strong conservation of site 1 compared to site 2 as expected.

3.3.1 XtractXact Filtered Results

The filtered_human_HYYHCAS.tsv file contained 48 potential SPS motif matches. These are results containing 25-35 base pair distance between the SPS motif match and a TATA box. The filtered_human_WYYMCAS.tsv contained 27 matches and the filtered_human_YTCHCAY.tsv file contained 7 matches (Table 18-20).

Chromosome	Strand	Gene	Gene Start	Gene Stop	Accession ID	Motif1	Spacer Length	Motif2	Motif Match	Motif Start	Motif Stop	Motif2TATA	TATA2Gene	Motif2Gene
chr10	+	C10ORF99	85933556	85945050	NM_207373.3	RTGRGAR	16	HYYHCAS	GTGAGAGAGATAGTCCTGTGCCCATTTCAC	85933025	85933054	472	29	503
chr8	-	CCNE2	95892452	95907429	NM_057749.3	RTGRGAR	15	HYYHCAS	ATGGGAGAATTCACTCTATTCCTTTCCAC	95909582	95909610	2124	29	2153
chr8	-	CCNE2	95892452	95907429	NM_057749.3	RTGRGAR	17	HYYHCAS	ATGGGAGAATTCACTCTATTCCTTTCCACAC	95909580	95909610	2122	29	2151
chr6	+	CFB	31913871	31919861	NM_001710.6_7	RTGRGAR	17	HYYHCAS	ATGAGAGTCCCTGCAAGGGTGAGTCCCTCAC	31912613	31912643	1198	29	1229
chr17	-	CHAD	48541857	48546241	NM_001267.3	RTGRGAR	16	HYYHCAS	GTGGGAGACCGACGCCGGGGCGGCTCTCAC	48546760	48546789	490	29	519
chr20	-	DEFB119	29964967	29978393	NM_153289.4	RTGRGAR	16	HYYHCAS	ATGGGAAGATTTGGAGAAAGAGCTTTACAG	29979353	29979382	932	28	960
chr20	-	DEFB119	29976771	29978393	NM_153323.5	RTGRGAR	16	HYYHCAS	ATGGGAAGATTTGGAGAAAGAGCTTTACAG	29979353	29979382	932	28	960
chr1	-	EVI5	92974253	93257967	NM_001308248.2	RTGRGAR	16	HYYHCAS	ATGGGAGGCCTTAGTTCTTTTCCTTCACAG	93260401	93260430	2403	31	2434
chr1	-	EVI5	92974253	93257967	NM_001308248.2	RTGRGAR	17	HYYHCAS	ATGAGAGGGCACAGGGCTTTTACTTTTCCAG	93259524	93259554	1526	31	1557
chr14	+	FOS	75745530	75748933	NM_005252.4	RTGRGAR	15	HYYHCAS	GTGAGAAAAAAAAGCGCCTAAATTCCCAC	75743565	75743593	1904	32	1938
chr11	+	FSHB	30252560	30256741	NM_001018080.3	RTGRGAR	16	HYYHCAS	GTGGGAGGGAAAGAGAGATACAGTTCACAC	30252081	30252110	420	29	451
chr11	+	GAL	68452011	68458643	NM_015973.5	RTGRGAR	15	HYYHCAS	GTGGGAGGGGAGGTCAGGCGAGTTTTCAG	68450837	68450865	1116	29	1147
chr12	+	GAPDH	6643682	6647537	NM_001357943.2	RTGRGAR	15	HYYHCAS	GTGGGAGGGGGCAGGGGACCTGTTCCCAC	6642553	6642581	1072	28	1102
chr6	-	GCLC	53362139	53409899	NM_001498.4	RTGRGAR	15	HYYHCAS	ATGGGAAGAACTCGTCCCAAGTCTCACAG	53410313	53410341	385	29	414
chr6	-	H2AC4	26033320	26033846	NM_003513.3	RTGRGAR	16	HYYHCAS	ATGGGAGAATTCCTCACAGCCTACCTCCAG	26033881	26033910	5	30	35
chr3	+	HES1	193853936	193856521	NM_005524.4	RTGRGAR	16	HYYHCAS	GTGGGAAAGAAAGTTTGGGAAGTTTCACAC	193853852	193853881	24	30	56
chr1	-	HES5	2460184	2461702	NM_001010926.4	RTGRGAR	16	HYYHCAS	GTGGGAACGGCCGCGGCGCCCGGACTCCAG	2461754	2461783	23	29	52
chr12	+	HIGD1AP9	50013049	50013458	HIGD1AP9	RTGRGAR	17	HYYHCAS	ATGGGAAGATAAAGTTGAGGAAATCTCCCAG	50012476	50012506	513	29	544
chr1	-	HRNR	152184552	152196669	NM_001009931.3	RTGRGAR	17	HYYHCAS	ATGGGAGAGTTTGACTAGATGATCTTTACAG	152197055	152197085	355	31	386
chr1	-	HSPB7	16340524	16344538	NM_001349689.2	RTGRGAR	16	HYYHCAS	ATGGGAAGAGATGTGAGGTCCCTTCCCCAG	16346647	16346676	2081	28	2109
chr5	+	IL13	131993864	131996802	NM_002188.3	RTGRGAR	16	HYYHCAS	GTGGGAGATGCCGTGGGCCCTCTACTACAG	131992359	131992388	1445	30	1477
chr17	-	KRT14	39738531	39743147	NM_000526.5	RTGRGAR	15	HYYHCAS	ATGGGAAAGTGTAGCCTGCAGGCCCACAC	39743365	39743393	189	29	218
chr17	-	KRT14	39738531	39743147	NM_000526.5	RTGRGAR	17	HYYHCAS	ATGGGAGGTCAGGCCTGGGAGGGCCCCCCAG	39743647	39743677	471	29	500
chr1	+	LCE2A	152670820	152671918	NM_178428.4	RTGRGAR	17	HYYHCAS	ATGGGAAGAAGCTGGAATAGCATACTTCCAC	152670472	152670502	288	29	319
chr20	+	MIR1-1HG	61147659	61167971	NM_001302812.2_2	RTGRGAR	17	HYYHCAS	GTGGGAAGCCGAGTGGCTATTTAGTTTCCAG	61146086	61146116	1510	32	1544
chr6	-	MIR4462	37523141	37523198	NR_039669.1	RTGRGAR	17	HYYHCAS	ATGAGAAAGAACGGTGACTCCGTCTCCCCAC	37523636	37523666	409	29	438
chr16	+	MT1B	56685797	56687116	NM_005947.3	RTGRGAR	15	HYYHCAS	GTGGGAAATCCAAAGGTCACAGCCTCCAC	56685509	56685537	230	29	261
chr16	-	MYH11	15796992	15950885	NM_001040113.2	RTGRGAR	16	HYYHCAS	GTGGGAGATGTCAAGTCAGATCCACCACAG	15951270	15951299	355	30	385
chr11	+	PATE1	125616173	125619762	NM_138294.3	RTGRGAR	15	HYYHCAS	GTGAGAAGGATCAGCTTACCTCTCTCCAG	125615511	125615539	604	29	635
chr11	-	PATE2	125646008	125648723	NM_212555.3	RTGRGAR	15	HYYHCAS	ATGGGAAGAACATTATTAATTCCTCCCAC	125649771	125649799	1014	34	1048
chr20	-	PIEZO1P1	58035674	58035892	PIEZO1P1	RTGRGAR	16	HYYHCAS	ATGGGAGCCACAGATGCTGCCAGCTTTCAG	58036542	58036571	622	28	650
chr13	-	POSTN	38136722	38172905	NM_006475.3	RTGRGAR	17	HYYHCAS	GTGGGAAAAGAACATGGATATGACATTACAG	38174028	38174058	1092	31	1123
chr6	-	PPP1R10	30568190	30585020	NM_001376195.1_7	RTGRGAR	17	HYYHCAS	ATGGGAACTGAGGACAAAGAATCTTCCTCAG	30587304	30587334	2252	32	2284
chr6	-	PPP1R10	30583330	30585020	NR_164781.1_7	RTGRGAR	17	HYYHCAS	ATGGGAACTGAGGACAAAGAATCTTCCTCAG	30587304	30587334	2252	32	2284
chr1	-	PRDX1	45976723	45987556	NM_002574.4	RTGRGAR	17	HYYHCAS	ATGGGAGGCTACATACTTTAAAGGTTTCCAG	45988314	45988344	728	30	758
chr12	-	PRR4	10998448	11002074	NM_007244.3_3	RTGRGAR	16	HYYHCAS	ATGAGAGGAAAGCAAAACCTAATCTTTCAG	11002156	11002185	55	27	82
chr8	+	PSCA	143761912	143764143	NM_005672.5_2	RTGRGAR	17	HYYHCAS	GTGGGAAGCTGATGTCCCCTGAACACCCCAG	143759475	143759505	2379	27	2408
chr9	-	RMRP	35657748	35658015	NR_003051.3	RTGRGAR	16	HYYHCAS	ATGGGAGATTCGGGCTCAGGGGATCCCCAC	35658777	35658806	730	32	762
chr3	-	SHOX2	157813694	157824209	NM_001163678.2	RTGRGAR	15	HYYHCAS	ATGGGAAAATTTATTGGTTTGAATTACAC	157826611	157826639	2372	30	2402
chr1	-	SLC2A5	9095165	9129694	NM_003039.3	RTGRGAR	16	HYYHCAS	ATGAGAAACTTTCCCCAAACAGTATTACAC	9130096	9130125	373	29	402
chr1	-	SLC2A5	9100139	9129694	NM_001135585.2	RTGRGAR	16	HYYHCAS	ATGAGAAACTTTCCCCAAACAGTATTACAC	9130096	9130125	373	29	402
chr4	+	SLC34A2	25657472	25680370	NM_001177998.2	RTGRGAR	15	HYYHCAS	GTGAGAACAGCCCCCTTCCTTTTCCTCAG	25656872	25656900	544	27	573
chr10	+	TECTB	114043157	114064796	NM_058222.3	RTGRGAR	17	HYYHCAS	ATGGGAGTAGCATGTTATCTGTTGTTTTCAG	114042480	114042510	617	29	648
chr6	-	TFAP2A	10396910	10419892	NM_001042425.3	RTGRGAR	17	HYYHCAS	ATGGGAAAGCAAACCCTAAACCAAACCTCAG	10420636	10420666	714	30	744
chr21	-	TFF1	43782391	43786644	NM_003225.3	RTGRGAR	15	HYYHCAS	ATGGGAAAGAGGGACTTTCTGAATCTCAG	43786857	43786885	184	29	213
chr14	-	TRK-TTT14-1	74055529	74056601	TRK-TTT14-1	RTGRGAR	15	HYYHCAS	ATGGGAACATCAGAAGGTACAATCTCCAG	74056497	74056525	869	27	896
chr6	+	TRT-AGT6-1	27130049	27130123	rna-TRT-AGT6-1	RTGRGAR	16	HYYHCAS	GTGGGAAGCAGGGTGACCACAGATCTCCAC	27129381	27129410	603	35	640
chr16	+	ZG16B	2880303	2882277	NM_145252.3	RTGRGAR	17	HYYHCAS	GTGGGAACCCAGCCCTCACCAGCCCCTCCAG	2878052	2878082	2192	28	2222

Table 18. Filtered results for 25-35bp distance between potential SPS motif and a potential TATA box for the HYYHCAS site 2 motif for the human promoter.

Chromosome	Strand	Gene	Gene Start	Gene Stop	Accession ID	Motif1	Spacer Length	Motif2	Motif Match	Motif Start	Motif Stop	Motif2TATA	TATA2Gene	Motif2Gene
chr8	-	CCNE2	95892452	95907429	NM_057749.3	RTGRGAR	15	WYYMCAS	ATGGGAGAATTCACTCTATTCCTTTCCAC	95909582	95909610	2124	29	2153
chr8	-	CCNE2	95892452	95907429	NM_057749.3	RTGRGAR	17	WYYMCAS	ATGGGAGAATTCACTCTATTCCTTTCCACAC	95909580	95909610	2122	29	2151
chr20	-	DEFB119	29964967	29978393	NM_153289.4	RTGRGAR	16	WYYMCAS	ATGGGAAGATTTGGAGAAAGAGCTTTACAG	29979353	29979382	932	28	960
chr20	-	DEFB119	29976771	29978393	NM_153323.5	RTGRGAR	16	WYYMCAS	ATGGGAAGATTTGGAGAAAGAGCTTTACAG	29979353	29979382	932	28	960
chr1	-	EVI5	92974253	93257967	NM_001308248.2	RTGRGAR	16	WYYMCAS	ATGGGAGGCCTTAGTTCTTTTCCTTCACAG	93260401	93260430	2403	31	2434
chr1	-	EVI5	92974253	93257967	NM_001308248.2	RTGRGAR	17	WYYMCAS	ATGAGAGGGCACAGGGCTTTTACTTTTCCAG	93259524	93259554	1526	31	1557
chr14	+	FOS	75745530	75748933	NM_005252.4	RTGRGAR	15	WYYMCAS	GTGAGAAAAAAAAGCGCCTAAATTCCCAC	75743565	75743593	1904	32	1938
chr11	+	FSHB	30252560	30256741	NM_001018080.3	RTGRGAR	16	WYYMCAS	GTGGGAGGGAAAGAGAGATACAGTTCACAC	30252081	30252110	420	29	451
chr12	+	GAPDH	6643682	6647537	NM_001357943.2	RTGRGAR	15	WYYMCAS	GTGGGAGGGGGCAGGGGACCTGTTCCCAC	6642553	6642581	1072	28	1102
chr3	+	HES1	193853936	193856521	NM_005524.4	RTGRGAR	16	WYYMCAS	GTGGGAAAGAAAGTTTGGGAAGTTTCACAC	193853852	193853881	24	30	56
chr1	-	HES5	2460184	2461702	NM_001010926.4	RTGRGAR	16	WYYMCAS	GTGGGAACGGCCGCGGCGCCCGGACTCCAG	2461754	2461783	23	29	52
chr1	-	HRNR	152184552	152196669	NM_001009931.3	RTGRGAR	17	WYYMCAS	ATGGGAGAGTTTGACTAGATGATCTTTACAG	152197055	152197085	355	31	386
chr1	-	HSPB7	16340524	16344538	NM_001349689.2	RTGRGAR	16	WYYMCAS	ATGGGAAGAGATGTGAGGTCCCTTCCCCAG	16346647	16346676	2081	28	2109
chr5	+	IL13	131993864	131996802	NM_002188.3	RTGRGAR	16	WYYMCAS	GTGGGAGATGCCGTGGGCCCTCTACTACAG	131992359	131992388	1445	30	1477
chr20	+	MIR1-1HG	61147659	61167971	NM_001302812.2_2	RTGRGAR	17	WYYMCAS	GTGGGAAGCCGAGTGGCTATTTAGTTTCCAG	61146086	61146116	1510	32	1544
chr6	-	MIR4462	37523141	37523198	NR_039669.1	RTGRGAR	17	WYYMCAS	ATGAGAAAGAACGGTGACTCCGTCTCCCCAC	37523636	37523666	409	29	438
chr16	-	MYH11	15796992	15950885	NM_001040113.2	RTGRGAR	16	WYYMCAS	GTGGGAGATGTCAAGTCAGATCCACCACAG	15951270	15951299	355	30	385
chr11	+	PATE1	125616173	125619762	NM_138294.3	RTGRGAR	15	WYYMCAS	GTGAGAAGGATCAGCTTACCTCTCTCCAG	125615511	125615539	604	29	635
chr13	-	POSTN	38136722	38172905	NM_006475.3	RTGRGAR	17	WYYMCAS	GTGGGAAAAGAACATGGATATGACATTACAG	38174028	38174058	1092	31	1123
chr1	-	PRDX1	45976723	45987556	NM_002574.4	RTGRGAR	17	WYYMCAS	ATGGGAGGCTACATACTTTAAAGGTTTCCAG	45988314	45988344	728	30	758
chr8	+	PSCA	143761912	143764143	NM_005672.5_2	RTGRGAR	17	WYYMCAS	GTGGGAAGCTGATGTCCCCTGAACACCCCAG	143759475	143759505	2379	27	2408
chr9	-	RMRP	35657748	35658015	NR_003051.3	RTGRGAR	16	WYYMCAS	ATGGGAGATTCGGGCTCAGGGGATCCCCAC	35658777	35658806	730	32	762
chr3	-	SHOX2	157813694	157824209	NM_001163678.2	RTGRGAR	15	WYYMCAS	ATGGGAAAATTTATTGGTTTGAATTACAC	157826611	157826639	2372	30	2402
chr1	-	SLC2A5	9095165	9129694	NM_003039.3	RTGRGAR	16	WYYMCAS	ATGAGAAACTTTCCCCAAACAGTATTACAC	9130096	9130125	373	29	402
chr1	-	SLC2A5	9100139	9129694	NM_001135585.2	RTGRGAR	16	WYYMCAS	ATGAGAAACTTTCCCCAAACAGTATTACAC	9130096	9130125	373	29	402
chr14	-	TRK-TTT14-1	74055529	74055601	TRK-TTT14-1	RTGRGAR	15	WYYMCAS	ATGGGAACATCAGAAGGTACAATCTCCAG	74056497	74056525	869	27	896
chr6	+	TRT-AGT6-1	27130049	27130123	rna-TRT-AGT6-1	RTGRGAR	16	WYYMCAS	GTGGGAAGCAGGGTGACCACAGATCTCCAC	27129381	27129410	603	35	640

Table 19. Filtered results for 25-35bp distance between potential SPS motif and a potential TATA box for the WYYMCAS site 2 motif for the human promoter.

Chromosome	Strand	Gene	Gene Start	Gene Stop	Accession ID	Motif1	Spacer Length	Motif2	Motif Match	Motif Start	Motif Stop	Motif2TATA	TATA2Gene	Motif2Gene
chr17	-	CHAD	48541857	48546241	NM_001267.3	RTGRGAR	16	YTCHCAY	GTGGGAGACCGACGCCGGGGCGGCTCTCAC	48546760	48546789	490	29	519
chr14	+	FOS	75745530	75748933	NM_005252.4	RTGRGAR	15	YTCHCAY	GTGAGAAAAAAAAGCGCCTAAATTCCCAC	75743565	75743593	1904	32	1938
chr11	+	FSHB	30252560	30256741	NM_001018080.3	RTGRGAR	16	YTCHCAY	GTGGGAGGGAAAGAGAGATACAGTTCACAC	30252081	30252110	420	29	451
chr12	+	GAPDH	6643682	6647537	NM_001357943.2	RTGRGAR	15	YTCHCAY	GTGGGAGGGGGCAGGGGACCTGTTCCCAC	6642553	6642581	1072	28	1102
chr3	+	HES1	193853936	193856521	NM_005524.4	RTGRGAR	16	YTCHCAY	GTGGGAAAGAAAGTTTGGGAAGTTTCACAC	193853852	193853881	24	30	56
chr9	-	LOC101927069	71568004	71590757	NR_110647.1	RTGRGAR	17	YTCHCAY	ATGAGAGGGATCTAGGTTTTGTGTTTCTCAT	71591624	71591654	835	32	867
chr11	-	PATE2	125646008	125648723	NM_212555.3	RTGRGAR	15	YTCHCAY	ATGGGAAGAACATTATTAATTCCTCCCAC	125649771	125649799	1014	34	1048

Table 20. Filtered results for 25-35bp distance between potential SPS motif and a potential TATA box for the YTCHCAY site 2 motif for the human promoter.

All unique genes found in each filtered file are shown in Table 21. There were forty-two genes found in the filtered_human_HYYHCAS.tsv file. Twenty-three genes were found in the filtered_human_WYYMCAS.tsv file and seven genes found in the filtered_human_YTCHCAY.tsv file.

HYYHCAS	WYYMCAS	YTCYCAY
C10ORF99	CCNE2	CHAD
CCNE2	DEFB119	FOS
CFB	EVI5	FSHB
CHAD	FOS	GAPDH
DEFB119	FSHB	HES1
EVI5	GAPDH	LOC101927069
FOS	HES1	PATE2
FSHB	HES5	
GAL	HRNR	
GAPDH	HSPB7	
GCLC	IL13	
H2AC4	MIR1-1HG	
HES1	MIR4462	
HES5	MYH11	
HIGD1AP9	PATE1	
HRNR	POSTN	
HSPB7	PRDX1	
IL13	PSCA	
KRT14	RMRP	
LCE2A	SHOX2	
MIR1-1HG	SLC2A5	
MIR4462	TRK-TTT14-1	
MT1B	TRT-AGT6-1	
MYH11		
PATE1		
PATE2		
PIEZO1P1		
POSTN		
PPP1R10		
PRDX1		
PRR4		
PSCA		
RMRP		
SHOX2		
SLC2A5		
SLC34A2		
TECTB		
TFAP2A		
TFF1		
TRK-TTT14-1		
TRT-AGT6-1		
ZG16B		

Table 21. Unique genes for each motif 2 for human promoter with a 25-35bp distance between potential SPS motif and a potential TATA box.

3.4 XtractXact and FIMO Performance Metrics Results

Figures 21 and 22 show the various performance metric results of XtractXact only. Figures 23 and 24 show performance metric results comparing XtractXact and FIMO.

XtractXact took approximately 4 minutes to run for the RTGRGAR[N*16]HYYHCAS motif for the intragenic region. It took approximately 40-60 seconds to run for the RTGRGAR[N*16]HYYHCAS motif on the various promoter regions (500, 1500, 2500bp). The intragenic region took longer due to the region being a larger search space than just the promoter regions of all human genes. In addition, as the search space gets larger for the promoter region (2500bp), the time it takes XtractXact to run multiple spacer lengths also increases.

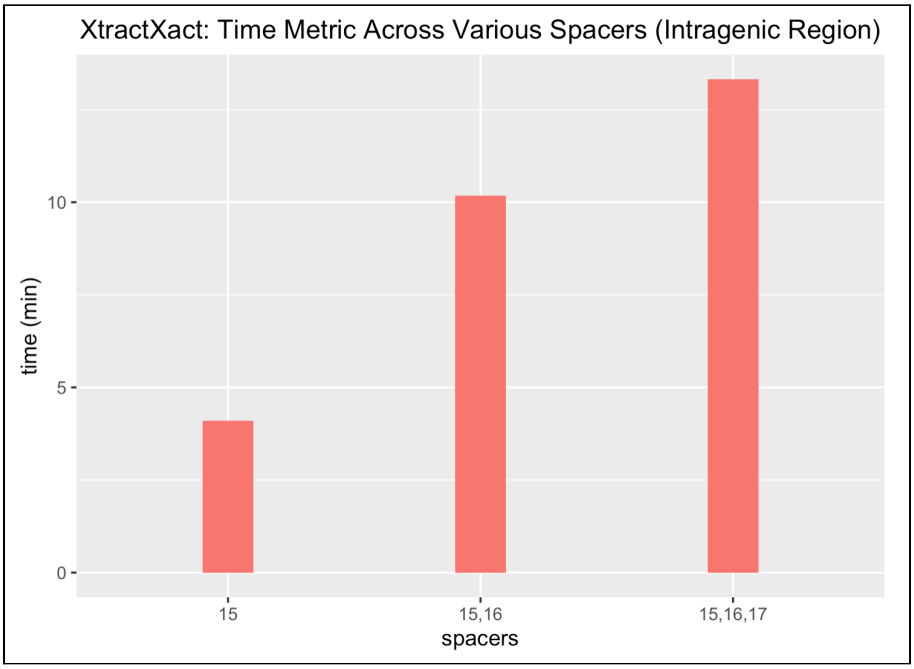


Figure 21. Time to run XtractXact with different spacer lengths for the RTGRGAR[N*16]HYYHCAS SPS motif within the intragenic regions.

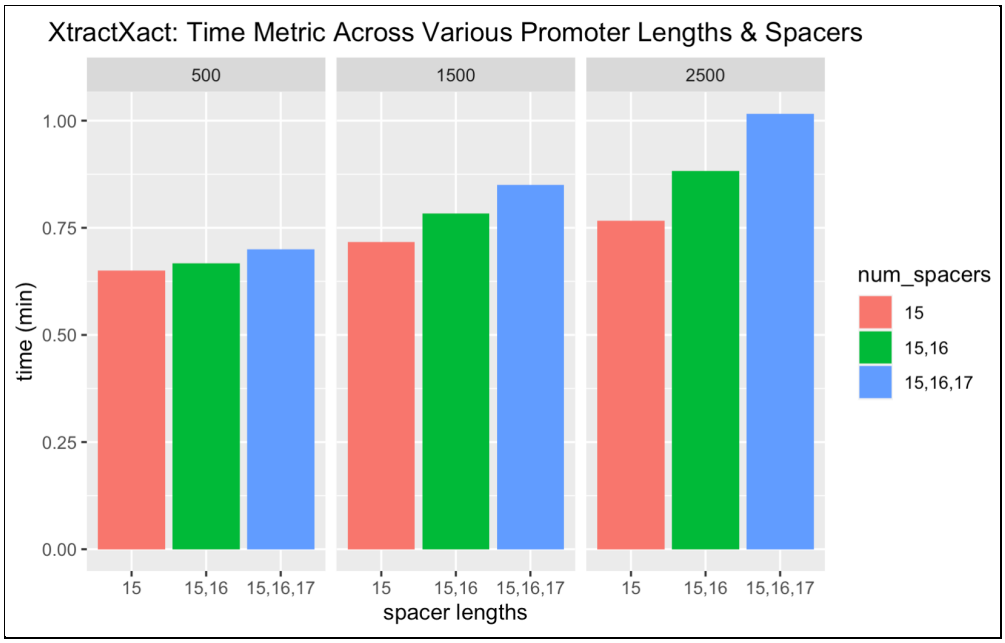


Figure 22. Time consumed to run XtractXact with the RTGRGAR[N*16]HYYHCAS SPS motif for different spacer lengths with the 500, 1500, and 2500 base pair promoter regions.

XtractXact ran approximately twice as fast as FIMO when using the same promoter and intragenic regions and number of motifs as shown in Figure 23 and 24. The larger the search space (2500bp promoter or intragenic regions), the longer it takes FIMO to report matches.

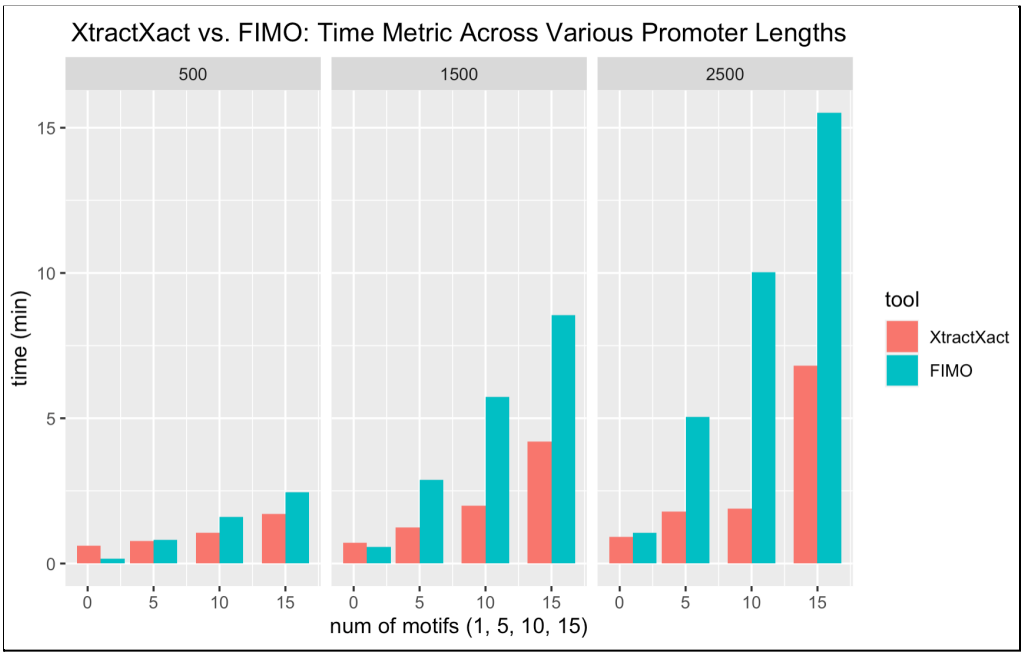


Figure 23. Time comparison between XtractXact and FIMO across various number of motifs (1, 5, 10, 15) and promoter lengths (500, 1500, 2500).

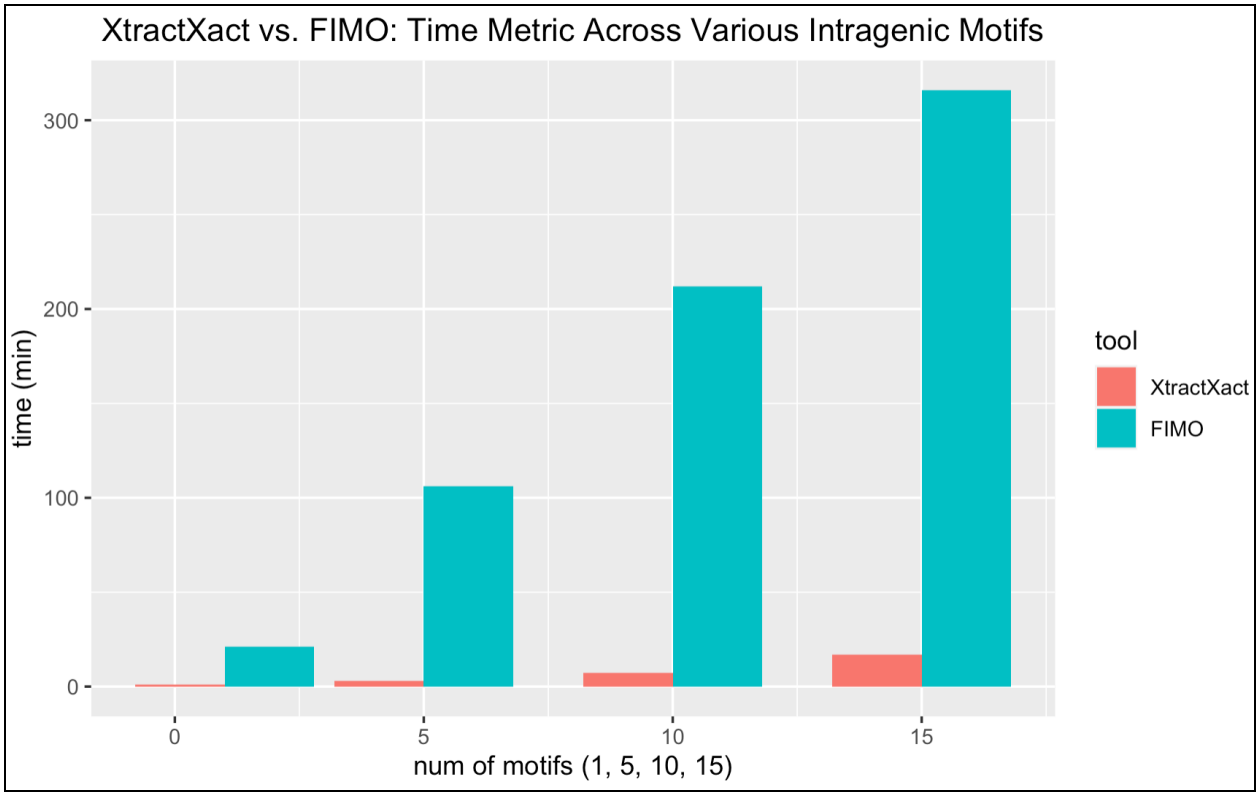


Figure 24. Time comparison between XtractXact and FIMO across various motifs for intragenic regions.

4. DISCUSSION

4.1 FIMO

FIMO can be used to extract potential sequences using a PWM. This PWM can be manipulated, but it requires the user to understand the underlying algorithms and calculations used to compute the p-value and scores by which the matches are reported. The results returned by FIMO include matches that do not align with the user's motif most likely due to the underlying PWM calculations. This results in sequence matches that include other nucleotides not provided by the user's motif at certain positions resulting in non-exact matches. The C nucleotides appear in the motif match results which do not correspond to the IUPAC code "D" for the DCYWSYS[N*16]MNKSGDA motif. Using different parameters to limit the number of potential mismatches for the user's motif is difficult to overcome. The UCSC genomic

coordinates given for the matched sequence results also do not match the true UCSC genomic coordinates found on the UCSC Genome Browser. These coordinates can be used to obtain gene information downstream of the motif match which cannot be done if the coordinates are incorrect.

By default, FIMO's web interface allows the user to upload or type a sequence database of interest in FASTA format. This option, however, is limited to 80MB which is approximately half the size of the 2500bp promoter region of all human genes (137MB). In addition, the results for a single motif of interest in the entire human genome is completed in approximately 30 minutes. However, when FIMO is run using a custom FASTA file with the intragenic regions only, the time increases dramatically as seen in Figure 24.

4.2 HMMER

HMMs are another useful statistical tool to train and build models to extract motifs from sequence databases. However, they are not ideal to use for very short motif sequences or motifs with long gapped regions as seen when using the RTGRGAR[N*16]HYYHCAS SPS motif or just the site1 or site 2 motifs like HYYHCAS. Short motifs frequently occur throughout the genome, so the probability calculations hmmer uses will not consider such short motifs as significant hits. Though profile HMMs allow for insertions and deletions, motifs containing long gapped regions are not ideal sequences to train an HMM.

4.3 XtractXact

XtractXact searches for exact motifs allowing the user flexibility through the various parameters available. In addition to finding motifs, XtractXact also provides the downstream gene and potential TATA box locations. This information will be useful to users who are seeking to identify functionally relevant motifs to study the effects of mutations on gene activation. In

addition to dimeric motif searches, XtractXact can potentially provide insight into trimeric or other more complex binding motifs if the sites are split prior to running the tool. The results from XtractXact can then be used to locate the matched motifs which are close in proximity to one another, potentially identifying more complex binding sites.

There currently exists no tool that is able to locate hundreds of motifs let alone millions or billions in a time efficient manner. By using job arrays, XtractXact can process large datasets in a time efficient manner.

4.4 XtractXact vs. FIMO

XtractXact and FIMO tools have their own advantages and disadvantages. FIMO runs slower than XtractXact due to the PWM calculations and the identification of more motifs which do not exactly match the user requested motif. The biggest performance difference between XtractXact and FIMO occurs when searching the intragenic regions. FIMO takes considerably longer than XtractXact possibly due to the large search space of the intragenic region compared to the promoter regions even when searching for a small number of motifs. Similarly, FIMO takes longer for longer upstream promoter lengths than XtractXact but performs better for less number of motifs searched (1 motif for the 500 base pair upstream region). FIMO is not ideal for high-throughput data containing hundreds to millions of motif searches. Though FIMO takes longer for the tested databases, FIMO provides the advantage of searching the entire human genome. However, further processing of the data needs to be performed to provide downstream genes and locations, TATA box locations, and other relevant information that might be useful for scientists.

4.5 Downstream Analyses

The intragenic region had more potential SPS results than the promoter region. The HYYHCAS motif in particular had the largest number of results at 2,078 matches for the promoter region and 20,416 matches for the intragenic region. This is due to the additional nucleotides using the IUPAC code H compared to the IUPAC code M and Y in the motifs WYYMCAS and YTCHCAY. YTCHCAY resulted in the least number of matches (251 for human promoter and 3,696 for human intragenic).

The web logo results in Tables 16 and 17 across the various motifs, spacer lengths, promoter, and intragenic regions show conserved nucleotides which match the site1 and site 2 SPS motifs. Figures 19 and 20 showing the *Hes1* and *Hes5* subtree web logos show similar site 1 and site 2 SPS conserved motifs. These subtrees contain slightly more conserved nucleotides at varying regions of the full length sequence due to using a smaller subset of the larger data.

Clustering using MEGAX used to create the web logos validated previous findings in the literature of potential cryptic sites for SPSs. The subtree for *Hes5* showed more conserved nucleotides in site 1 than the subtree for *Hes1* site 1, specifically the GGG nucleotides in site 1 of the *Hes5* subtree weblogo. This shows that cryptic sites such as the site 2 for *Hes5* might require strongly conserved site 1 nucleotides that bind to CSL containing the GGG nucleotides. The full dataset web logos show no conserved nucleotides within the spacer region, but the *Hes1* and *Hes5* subtrees show some conservation at these positions. If the *Hes1* and *Hes5* subtrees include true SPSs, these conserved nucleotides should be further investigated for importance to other transcriptional elemental contacts.

4.6 Future Directions

Though XtractXact provides exact motif matches and gene information, further improvements can be made. In addition to motif searching and gene finding, XtractXact can be expanded to include searches for other transcription factor binding sites near the matched motif sites using the JASPAR PWM database. This can provide another additional tool to identify regulatory elements surrounding motifs of interest. Currently, XtractXact automatically iterates through input site 1 and site 2 motif files such that each site 1 is paired with every site 2 motif. XtractXact can be modified to allow pairing of row 1 in the site 1 motif file with pair 1 in the site 2 motif file. This can prevent unwanted site 1 and site 2 combinations.

XtractXact allows for promoter and intragenic searches only. This is due to the time it takes the tool to find exact motif matches for an entire human genome. To allow for entire genome searches, Python multithreading can be incorporated to quicken tool performance. If the results from FIMO can be used to obtain initial results for whole genome search, it can be potentially extended to include relevant gene and motif information that XtractXact provides if the UCSC genomic coordinates are correct. Combining both tools to take advantage of what each has to offer can improve performance for motif searches in genomes.

A difficulty that arises when comparing XtractXact results from the human genome to the mouse genome is the consistency of the gene names. UCSC Genome Browser follows the RefSeq gene naming scheme and might not easily convert to gene names used in other databases. Instead of using RefSeq, future work will use Ensembl gene IDs as these IDs can be found across various databases. If this can be overcome, the results across species can be further analyzed to find potential overlapping genes and SPSs which can be used to perform gene analyses.

The website version of XtractXact is currently in development. This will allow users to readily use the tool without learning how to run tools via the command line making it more accessible. The website version requires the same parameters as the command line version, and future developments will allow for the exploration of the results for phylogenetic tree building and web logos.

REFERENCES

[1] Mumm, J. S., Kopan, R., “Notch signaling: From the outside,” Dev. Biol., vol. 228, no. 2, p. 151-165, Dec. 15, 2000.

[2] H. Liu, A. W. S. Chi, K. L. Arnett, M. Y. Chiang, L. Xu, O. Shestova, et al., “Notch dimerization is required for leukemogenesis and T-cell development,” Genes Dev., vol. 24, pp. 2402, October 2010.

[3] Ramos C.A., "MASTERMIND-LIKE 1-Dependent Notch Target Gene Activation Requires a Sequence-Paired Site and a TATA Box", Master's Theses, 4360, 2013, https://scholarworks.sjsu.edu/etd_theses/4360.

[4] K. Arnett, M. Hass, D. McArthur, et al., “Structural and mechanistic insights into cooperative assembly of dimeric Notch transcription complexes,” Nat. Struc. & Mol. Bio., vol. 17, no.11, p. 1312-1317, Nov. 2010.

[5] Y. Yashiro-Ohtani, H. Wang, C. Zang, K. Arnett, et al., “Long-range enhancer activity determines Myc sensitivity to Notch inhibitors in T cell leukemia,” PNAS, p. E4946-E4953, Nov. 4, 2014.

[6] T. Tun, Y. Hamaguchi, N. Matsunami, et al., “Recognition sequence of a highly conserved DNA binding protein RBP-Jx,” Nuc. Acids. Res., vol. 22, no. 6, p.965-971, March 1994.

[7] Fryer C. J., Lamar E., Turbachova I., et al., “Mastermind mediates chromatin-specific transcription and turnover of the Notch enhancer complex,” Genes and Dev., vol. 16, no. 11, p.1397-1411, 2002.

[8] J. Crow, A. Albig, “Notch family members follow stringent requirements for intracellular domain dimerization at sequence-paired sites,” PLOS One, vol.15, no.11, p.1-19, Nov. 24, 2020.

[9] Ong C.T., Cheng H.T., Chang L.W., et al., “Target Selectivity of Vertebrate Notch Proteins: Collaboration Between Discrete Domains and CSL-Binding Site Architecture Determines Activation Probability,” JBC, vol. 281, no. 8, p. 5106-5119, 2006.

[10] L. Elnitski, V. X. Jin, P. J. Farnham, and S. J. Jones, “Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques,” Genome Res., vol. 16, pp.1455-1464, December 2006.

[11] Riley, T., Sontag, E., Chen, P. *et al.*, “Transcriptional control of human p53-regulated genes,” Nat. Rev. Mol. Cell Biol., vol. 9, p. 402–412, May 2008.

[12] Narlikar, L., Ivan O., “Identifying Regulatory Elements in Eukaryotic Genomes,” Briefings in Functional Genomics & Proteomics, vol. 8, no. 4, p. 215-230, July 2009.

[13] Staden, R., “Computer methods to locate signals in nucleic acid sequences,” Nucleic Acids Res., vol. 12, p. 505-519, Jan. 11, 1984.

[14] Grant C.E., Bailey T.L., Noble W.S., “FIMO: scanning for occurrences of a given motif,” J. Bioinform., vol. 27, no. 7, p. 1017-1018, Apr. 1, 2011.

[15] Baldi P., Chauvin Y., Hunkapiller T., McClure M.A., “Hidden Markov models of biological primary sequence information,” Proc. Natl. Acad. Sci. U.S.A., vol. 91, no. 3, p. 1059-1063, Feb. 1, 1994.

[16] Jurafsky D., Martin J.H., “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,” Upper Saddle River, N.J., 2000.

[17] Embl-Ebi, “What Are Profile Hidden Markov Models," EMBL-EBI Training Programme.

[18] Sievers F., Higgins D.G., “Clustal Omega for making accurate alignments of many protein sequences,” Protein Sci., vol. 27, p. 135-145, Jan. 2018.

[19] Wheeler T.J., Eddy S.R., “nhmmer: DNA Homology Search With Profile HMMs,” J. Bioinform., vol. 29, p. 2487-2489, 2013.

[20] Johnson L.S., Eddy S.R., Portugaly E., “Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure,” BMC Bioinform., vol. 11, p. 431, 2010.

[21] M. Berger and M. Bulyk, “Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors,” Nat. Protoc., vol. 4, pp.393–411, March 2009.

[22] C. D. Bianco, A. Vedenko, S. H. Choi, et al., “Notch and MAML-1 complexation do not detectably alter the DNA binding specificity of the transcription factor CSL,” PLoS One, vol. 5, pp. e15034, November 2010.

[23] E. Severson, K. L. Arnett, H. Wang, C. Zang, L. Taing, H. Liu, et al., “Genome-wide identification and characterization of Notch transcription complex-binding sequence-paired sites in leukemia cells,” Sci. Signal., vol. 1, pp. 1–10, May 2017.

[24] Baum D., “Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups,” Nat. Sci. Educ., vol. 1, p. 190, 2008.

[25] Russo C.A.M., Selvatti A.P., “Bootstrap and Rogue Identification Tests for Phylogenetic Analyses,” Mol. Biol. Evol., vol. 35, no. 9, p. 2327-2333, June 7, 2018.

[26] Crooks G.E., Hon G., Chandonia J.M., et al., “WebLogo: A sequence logo generator,” Genome Res., vol. 14, p. 1188-1190, 2004.

[27] S.R. Eddy, “Profile hidden Markov models,” Bioinformatics, vol. 14, no. 9, p. 755-763, Oct. 1, 1998.

[28] Stecher G., Tamura K., Kumar S., “Molecular Evolutionary Genetics Analysis (MEGA) for macOS,” Mol. Biol. Evol., vol. 37, no. 4, p. 1237-1239, Apr. 2020.

[29] Sneath P. H. A., “Numerical taxonomy: the principles and practice of numerical classification,”.J. Gen. Microbiol., vol. 17, p. 201-226. 1957.