reproductive medicine

IMPACT FACTOR 1.1

Article

Exploring Regulatory Properties of Genes Associated with Nonsyndromic Male Infertility

Daniela Hristov and Done Stojanov

MDPI

https://doi.org/10.3390/reprodmed5030013 reproductive medicine

MDPI

Article

Exploring Regulatory Properties of Genes Associated with Nonsyndromic Male Infertility

Daniela Hristov [1,*] and Done Stojanov [2,*]

[1] IVF Laboratory, Re-Medika General Hospital, 1000 Skopje, North Macedonia
[2] Faculty of Computer Science, Goce Delcev University, 2000 Stip, North Macedonia
[*] Correspondence: dhristov@remedika.com.mk (D.H.); done.stojanov@ugd.edu.mk (D.S.)

Abstract: In this study, we analyzed the regulatory properties of 26 (twenty-six) genes associated with nonsyndromic male infertility. We applied an in silico analysis in order to determine the number and distribution of promoters and identify relevant promoter consensus sequences and potential transcription factors. Underlining the concept of alternative transcriptional initiation (ATI), we have found that 65.4% of genes associated with nonsyndromic male infertility have 1 (one) to 6 (six) promoters, located in the region 1 kb upstream of the TSS, and 41% of them are located at a position below −500 bp. Although the TATA box consensus sequence TAWAAA, such as W is A or T, appears at a common location in all genes, it is shifted for at least 10 bp in the EFCAB9 gene. The C2H2 zinc finger is found to be the most significant common transcription factor, binding genes' promoters GLIS1, ZSCAN21, GLIS3, GLIS1, ZNF770, ZNF780A, ZNF81, and ZNF264. On the other hand, basic leucine zipper factors (bZIPs) bind the JUNB gene promoter specifically, exhibiting unique regulatory properties of all genes associated with nonsyndromic male infertility. Two genes, NANOS1 and ZMYND15, are expected to be less susceptible to DNA methylation, due to the high density of CpG content found in their promoter regions.

Keywords: nonsyndromic male infertility; genes; promoters; consensus sequence; TATA box; transcription factors; CpG islands; in silico analysis

check for updates

Citation: Hristov, D.; Stojanov, D. Exploring Regulatory Properties of Genes Associated with Nonsyndromic Male Infertility. *Reprod. Med.* **2024**, *5*, 136–153. https://doi.org/10.3390/reprodmed5030013

Academic Editor: Berthold Huppertz

Received: 2 April 2024
Revised: 24 July 2024
Accepted: 26 July 2024
Published: 2 August 2024

1. Introduction

Infertility, a term for the inability of organisms to naturally propagate, involves a complex interaction of molecular, hormonal, and genetic pathways, particularly notable in the context of human reproduction. Defined by the failure to conceive or sustain a viable pregnancy after a year of regular controlled ovulation and unprotected sexual intercourse [1], infertility affects a substantial portion of the global population, with approximately 48.5 million couples, constituting 15% of the couples worldwide [2]. A healthy young couple typically faces a modest 20–25% chance of conception per menstrual cycle, highlighting the multifactorial nature of conception [3]. Factors such as hormonal imbalances, age-related declines, lifestyle influences (e.g., physical activity, obesity), infectious diseases, immunological factors, psychological stressors, surgical interventions, and anatomical obstructions contribute to infertility, often with underlying genetic predispositions [3].

Notably, genetics plays a significant role in male infertility, accounting for 15–30% of cases [4,5]. However, infertility is not associated with a single gene, but a lot of chromosomal aberrations, single-gene mutations, and multifactorial inheritance patterns together contribute to its etiology. Chromosomal abnormalities and single-gene mutations, for instance, encompass about 10–15% of male infertility cases [6]. Male infertility, comprising 50% of all infertility cases, presents a complex clinical landscape, with causative factors remaining unidentified in 30% of cases [7].

The complex nature of the process explains why the causes of infertility are identified in only a portion of cases. Indeed, about 40% of cases remain undiagnosed and are classified as "idiopathic" [8]. It is believed that approximately 50% of these idiopathic cases could be due to genetic defects [9].

Since the 1970s, it has been known that genetic anomalies can affect human fertility [10]. These anomalies fall into two categories: (i) karyotype anomalies, involving numerical or structural changes, and (ii) genetic anomalies, affecting a single specific gene. When genetic anomalies are present, syndromic conditions can be identified, where infertility is one of several symptoms linked to a pathological syndrome This is known as syndromic infertility, where infertility is usually not the primary issue. In contrast, nonsyndromic infertility is caused by gene mutations that lead to absent or abnormal spermatogenesis, without any other symptoms. Recent progress in molecular biology and medical genetics has facilitated the discovery of the genetic causes of male infertility. Over the past decade, a new research domain referred to as the "genetics of infertility" has emerged. Recent advances in biocomputing [11–13] and whole genome sequencing techniques have allowed the identification of an increasing number of gene mutations responsible for specific infertility phenotypes. Early achievements in this area have attracted numerous new researchers. We anticipate that the list of infertility-associated genes will significantly grow over the next decade, leading to more available diagnostic tests. As a result, the number of idiopathic infertility cases is expected to decline, as diagnostic testing will be available to more couples. This research aims to explore the regulatory properties of genes associated with infertility phenotypes [14].

Promoters, short DNA regions (100–1000 bp) located proximal to transcription start sites (TSSs), regulate gene expression. Promoters control DNA transcription by direct interaction with basal transcription machinery components, such as RNA Polymerase II and transcription factors. They can be classified as core, proximal, or distal, depending on the promoter location relative to the TSS [15]. Identifying promoters is the key to defining transcription units, decoding gene structure, uncovering regulatory mechanisms, and annotating gene function [16]. Within promoter regions, conserved DNA motifs are crucial for gene regulation, and their systematic identification enhances our understanding of regulatory networks [17].

Transcription factors (TFs) are regulatory proteins whose function is to activate (or more rarely, to inhibit) the transcription of DNA by binding to specific DNA sequences [18]. TFs have defined DNA-binding domains, with an up to 106-fold higher affinity for their target sequences than the rest of the DNA strand. These highly conserved sequences have been used to categorize the known TFs into various "families" [19].

The TATA box is recognized in a sequence-specific manner by the TATA box-binding protein (TBP), an essential factor involved in the initiation of transcription by all three eukaryotic RNA polymerases. The TATA box sequence in eukaryotes is located about 25 bp upstream of many genes transcribed by RNA polymerase II (Pol II) and some genes transcribed by RNA polymerase III (Pol III). The TATA box was originally identified as a regulatory signal upstream of many protein-coding genes transcribed by RNA polymerase II (Pol II) [19]. However, some tRNA and 5S RNA genes and most RNA polymerase III (Pol III)-transcribed genes with external promoters also contain TATA boxes 25–30 bp upstream of the transcription start site. When present in Pol III promoters, the TATA box can have a significant effect on the efficiency and accuracy of the transcription of these genes by Pol III [20–23].

CpG islands (CGIs) are genomic regions containing a high density of CpG dinucleotide repeats. In mammalian genomes, CpG islands typically span 300–3000 base pairs and are commonly found within or near approximately 40% of gene promoters. Importantly, CpG dinucleotides within CpG islands are often unmethylated, especially in regions rich in GC pairs like CpG clusters and CpG islands, which is a key feature of gene promoters and gene expression control. The hypermethylation of CpG islands near promoters is associated with the transcriptional silencing of the corresponding genes. DNA methylation induces gene silencing through various mechanisms, including the inhibition of transcription factor binding and the alteration of chromatin structure, which can directly impact histone acetylation and regulate the higher-order chromatin structure.

Numerous studies have demonstrated that promoter hypermethylation can lead to the downregulation of key genes involved in various signaling pathways, such as cell cycle regulation, apoptosis, DNA repair, drug resistance, detoxification, angiogenesis, invasion, and metastasis [24].

In silico applications have emerged as promising resources in biological research, assisting the rapid extraction of meaningful insights from biological data and driving advances in bioinformatics and computational biology [25–28]. According to previous studies [29–32], a lot of genes are associated with nonsyndromic male infertility.

This study aims to analyze the regulatory properties of 26 (twenty-six) genes associated with nonsyndromic male infertility in Yahaya et al., 2020 [29]. With the aid of in silico applications, we aim to analyze the promoters' structure, identify alternative transcriptional initiation sites (ATIs), analyze common motifs, and identify potential transcription factors. We aim to contribute a deeper understanding of the genetic underpinnings of male unexplained infertility.

2. Materials and Methods

2.1. Determination of Promoter Regions for Genes Associated with Nonsyndromic Male Infertility

Twenty-six sequences encoding genes associated with nonsyndromic male infertility in [29] were retrieved in FASTA format from the National Center for Biotechnology Information (NCBI) Genome Browser "https://www.ncbi.nlm.nih.gov/gene (accessed on 1 March 2024)" and in silico analyzed. Genes associated with syndromic infertility were excluded from the scope of our analysis. The promotor region of each gene was identified using the online Neural Network Promoter Prediction (NNPP version 2.2) application (BDGP: Neural Network Promoter Prediction) "https://fruitfly.org/seq_tools/promoter.html (accessed on 1 March 2024)". According to prior methodologies [33–36], a minimum of 1 kilobase (kb) upstream of the gene's known transcription start site (TSS) needs to be considered in order to identify gene promoters. The retrieved records were analyzed with NNPP v.2.2, with a cut off value of 0.8 for significant promoter predictions [37]. Although there is no strict promoter predictivity score threshold-level limitation and it could be set up either higher or lower, other studies, such as [38], employ the same threshold level. Following NNPP (Neural Network Promoter Prediction) program cross-validation results on a dataset of unrelated eukaryotic genes, the number of false positives is expected to range between 0.4 and 0.8%, given the threshold level of 0.8, which is a tolerated rate of error in addition to our analysis. In a case with multiple promoters, the prediction of the promoter with the highest predictive score was considered as statistically most significant [39]. Promoter regions of interest were 1 kb regions upstream of the known TSS of each gene.

2.2. Determination of Common Motifs and Transcription Factors for Promoter Regions of Genes Associated with Nonsyndromic Male Infertility

We used the web-based analysis program MEME (Multiple Em for Motif Elicitation; version 5.5.5: "https://meme-suite.org/meme/tools/meme (accessed on 1 March 2024)" to search for the common motifs within the identified promoters of the genes associated with nonsyndromic male infertility [40]. The motifs' lengths ranged between 6 and 50 bp, searching up to 5 (five) motifs. The resulting MEME output, in HTML format, containing significant consensus motifs, was then parsed to the TOMTOM (Motif Comparison Tool) "https://meme-suite.org/meme/doc/tomtom-output-format.html (accessed on 1 March 2024)" [41] web server for the identification of likely transcription factors (TFs) binding the identified motifs. TOMTOM operates by comparing one or more motifs against a database of known motifs, ranking them accordingly, and generating alignments for each significant match.

2.3. Gene Ontology Analysis

We used the GOMo (Gene Ontology for Motif) application "https://meme-suite.org/meme/doc/gomo.html?man_type=web (accessed on 1 March 2024)" [42] to scan known promoters against nucleotide motifs identified by the MEME application. This analysis aimed to determine if any motif exhibited a significant association with the genes linked to one or more Genome Ontology (GO) terms, suggesting the biological roles of the motifs if significant GO terms identified. GOMo operates by searching through a set of ranked genes to identify enriched GO terms that are associated with high-ranking genes.

2.4. Search for CpG Islands

We used the database of CG-rich islands and analytical tool (DBCAT) "http://dbcat.cgm.ntu.edu.tw/ (accessed on 1 March 2024)" to search for CpG islands. This program applies string-processing methods to detect CpG islands, based on the criterion CG content $\geq$ 55%, Observed CpG/Expected CpG ratio $\geq$ 0.65, and length $\geq$ 500 bp [43].

3. Results

3.1. Identification of Promoters

Promoters were predicted for each of the 26 (twenty-six) genes associated with nonsyndromic male infertility predisposition (Table 1). The NNPP application did not identified any known promoter for the genes SPATA, AURC, CATSPER, SYCP3, SYCP2, DAZ1, XRCC2, TEX11, and TAF4BF, located in the region 1 kb upstream of the TSS, excluding them from further analysis.

Table 1. Predictive score and number of promoters for each gene associated with nonsyndromic male infertility.

Gene Symbol (Gene ID) Full Name *	Corresponding Promoter Region Name	No. of Promoters Identified in Promoter Region (1000 bp Upstream)	Predictive Score at Cut Off Value 0.8 **	Distance from Start Codon (ATG) to Upstream
SPATA16 (ID: 83893) spermatogenesis associated 16 gene	Prom_SPATA	0		
AURKC (ID: 6795) aurora kinase C	Prom_AURC	0		
CATSPER1 (ID: 117144) cation channel sperm associated 1	Prom_CATSPER	0		
MTHFR (ID: 4524) methylenetetrahydrofolate reductase	Prom_MTHFR	2	0.87, 0.98	−3719, −3709
EFCAB9 (ID: 285588) EF-hand calcium binding domain 9	Prom_EFCAB9	2	0.97,0.96	−738, −305
FKBP6(ID: 8468) FKBP prolyl isomerase family member 6 (inactive)	Prom_ FKBP6	2	0.85, 0.82	−1181, −484
SYCP3(ID: 50511) synaptonemal complex protein 3	Prom_SYCP3	0		
HSF2 (ID: 3298) heat shock transcription factor 2	Prom_HSF2	3	0.95, 0.85, 0.99	−720, −622, −328
SYCP2 (ID: 10388) synaptonemal complex protein 2	Prom_SYCP2	0		
MYBL1 (ID: 4603) MYB proto-oncogene like 1	Prom_MYBL1	2	0.86, 1.00	−989, −966
KIT (ID: 3815) KIT proto-oncogene, receptor tyrosine kinase	Prom_KIT	2	0.92, 0.82	−772, −203

Table 1. *Cont.*

Gene Symbol (Gene ID) Full Name *	Corresponding Promoter Region Name	No. of Promoters Identified in Promoter Region (1000 bp Upstream)	Predictive Score at Cut Off Value 0.8 **	Distance from Start Codon (ATG) to Upstream
KLHL10 (ID: 317719) kelch like family member 10	Prom_KLHL10	1	0.96	−2604
NANOS1 (ID: 340719) nanos C2HC-type zinc finger 1	Prom_NANOS1	6	0.96, 0.85, 0.80, 0.85, 1.00, 0.87	−611, −588, −234, −95, −86, −65
PRM1 (ID: 5619) protamine 1	Prom_PRM1	4	0.84, 0.98, 0.93, 1.00	−736, −355, −101, −92
PRM2 (ID: 5620) protamine 2	Prom_PRM2	3	0.80, 0.93, 1.00	−319, −113, −105
SEPTIN12 (ID: 124404) septin 12	Prom_SEPT12	3	1.00, 0.97, 0.92	−1226, −1083, −579
TNP1 (ID: 7141) transition protein 1	Prom_TNP1	3	0.93, 0.96, 0.93	−973, −192, −33
TNP2 (ID: 7142) transition protein 2	Prom_TNP2	1	0.99	−61
DAZ1(ID: 1617) deleted in azoospermia 1	Prom_DAZ1	0		
XRCC2 (ID: 7516) X-ray repair cross complementing 2	Prom_ XRCC2	0		
ZMYND15 (ID: 84225) zinc finger MYND-type containing 15	Prom_ZMYND15	1	0.90	−736
TEX11 (ID: 56159) testis expressed 11	Prom_TEX11	0		
ADGRG2 (ID: 10149) adhesion G protein-coupled receptor G2	Prom_ ADGRG2	5	0.80, 0.84, 0.99, 0.99, 0.99	−54850, −54799, −54573, −54547, −54319
CCDC62 (ID: 84660) coiled-coil domain containing 62	Prom_ CCDC62	1	0.87	−270
TAF4B (ID: 6875) TATA-box binding protein associated factor 4b	Prom_TAF4B	0		
GALNTL5 (ID: 168391) polypeptide N-acetylgalactosaminyltransferase like 5	Prom_GALNTL5	3	0.99, 0.97, 0.92	−11388, −11254, −11021

* provided by HGNC "https://www.genenames.org/ (accessed on 1 March 2024)" ** Cut off value is set to 0.8 for reliable predictions.

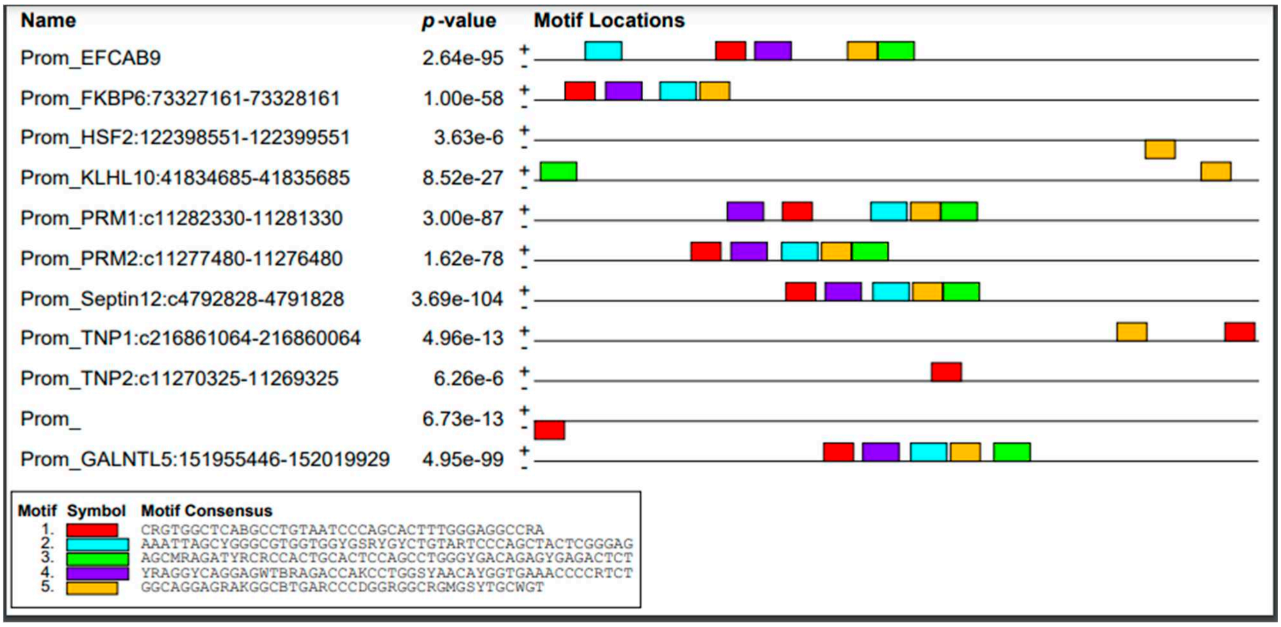
We have identified a single promoter sequence in the region of interest for the genes KLHL10, TNP2, ZMYND15, and CCDC62 (Table 1), while all others displayed multiple promoters (ranging from 2 (two) to 6 (six)), Table 1. The predicted promoters for all genes had a predictive score ranging from 0.80 to 1.00. The majority of the predicted promoters (41%) were located below −500 bp upstream from the start codon (ATG)—1 (one) promoter: EFCAB9, FKBP6, HSF2, KIT, TNP2, CCDC 62, 2 (two) promoters: TNP1, 3 (three) promoters: PRM1 and PRM2, and 4 (four) promoters: NANOS1 (Table 1). The rest were distributed between −500 and −3000 bp (36%) (1 (one) promoter: EFCAB9, FKBP6, KIT, KLHL10, PRM1, TNP1, ZMYND15, 2 (two) promoters: HSF2, MYBL1, NANOS1, 3 (three) promoters: SEPTIN) and beyond −3000 bp (23%) (2 (two) promoters: MTHFR, 3 (three) promoters: GALANTL5, 5 (five) promoters: ADGRG), Table 1.

3.2. Common Candidate Motifs and Transcription Factors of Genes Associated with Nonsyndromic Male Infertility

We used the motif-based sequence analysis tool MEME "https://meme-suite.org/meme/ (accessed on 1 March 2024)" [40] to search for common motifs within the identified promoters. Five candidate motifs were discovered by the MEME algorithm (Table 2). The MEME application generated common candidate motifs for 14 (fourteen) nonsyndromic male infertility-associated genes' promoters. The identified motifs are distributed on both positive (34 (thirty-four)) and negative (2 (two)) strands. The majority of candidate common motifs in the promoter regions are densely located between −500 and −1000 bp of the TSSs (Figure 1). Motifs that were shared by the majority of the promoter regions of the genes associated with nonsyndromic male infertility were chosen for the determination of a functionally important candidate motif. The number of binding sites within the identified common motifs ranged between 6 (six) and 9 (nine), Table 2. The common motifs' length was 41 or 50 bp, Table 2. Two motifs, Motif1 and Motif5, were found in 64.2% of promoters, while 3 (three) of the common motifs, Motif [2–4], were shared among 42.8% of promoters. The motif with the highest e-value, being also common for 64.2% of the input sequences, was Motif1, Table 2, Figures 1 and 2.

Table 2. Promoters' common motifs (1 kb upstream of the TSS).

Discovered Candidate Motifs	Number (%) Promoters Containing Each One of the Motifs	E-Value	Motif Width	Number of Binding Sites
Motif1	9 (64.2%)	1.20e-43	41	9
Motif2	6 (42.8%)	2.00e-44	50	6
Motif3	6 (42.8%)	2.00e-38	50	6
Motif4	6 (42.8%)	1.40e-33	50	6
Motif5	9 (64.2%)	1.30e-18	41	9



Name	*p*-value
Prom_EFCAB9	2.64e-95
Prom_FKBP6:73327161-73328161	1.00e-58
Prom_HSF2:122398551-122399551	3.63e-6
Prom_KLHL10:41834685-41835685	8.52e-27
Prom_PRM1:c11282330-11281330	3.00e-87
Prom_PRM2:c11277480-11276480	1.62e-78
Prom_Septin12:c4792828-4791828	3.69e-104
Prom_TNP1:c216861064-216860064	4.96e-13
Prom_TNP2:c11270325-11269325	6.26e-6
Prom_	6.73e-13
Prom_GALNTL5:151955446-152019929	4.95e-99

Motif	Symbol	Motif Consensus
1.	🟥	CRGTGGCTCABGCCTGTAATCCCAGCACTTTGGGAGGCCRA
2.	🟦	AAATTAGCYGGGCGTGGTGGYGSRYGYCTGTARTCCCAGCTACTCGGGAG
3.	🟩	AGCMRAGATYRCRCCACTGCACTCCAGCCTGGGYGACAGAGYGAGACTCT
4.	🟪	YRAGGYCAGGAGWTBRAGACCAKCCTGGSYAACAYGGTGAAACCCCRTCT
5.	🟧	GGCAGGAGRAKGGCBTGARCCCDGGRGGCRGMGSYTGCWGT

Figure 1. Block diagrams showing the distribution and location of the candidate common motifs in different genes associated with nonsyndromic male infertility, upstream of the TSSs, represented with their symbols.

Motif1's sequence is "CRGTGGCTCABGCCTGTAATCCCAGCACTTTGGGAGGCCRA" (Figure 2), such as, except for the four DNA nucleotides, R stands for Guanine (G) and Adenine (A) found at an equal frequency, and B stands for an equal distribution of Guanine (G), Cytosine (C), and Thymine (T) in the motif's frequency matrix [41]. Nucleotides with a constant presence in all 9 (nine) genes containing Motif1 in their promoters are Guanine on positions 6 and 34, Cytosine on positions 14 and 28, and Adenine on positions 18, 19, and 27.

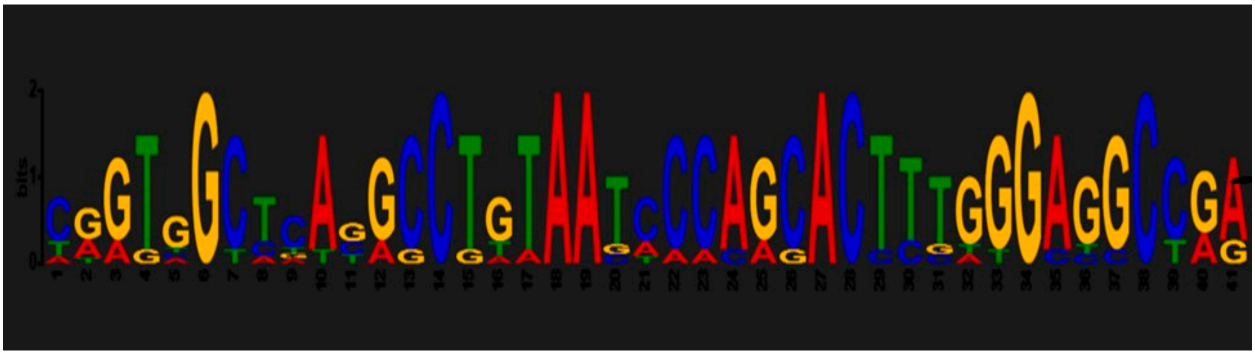
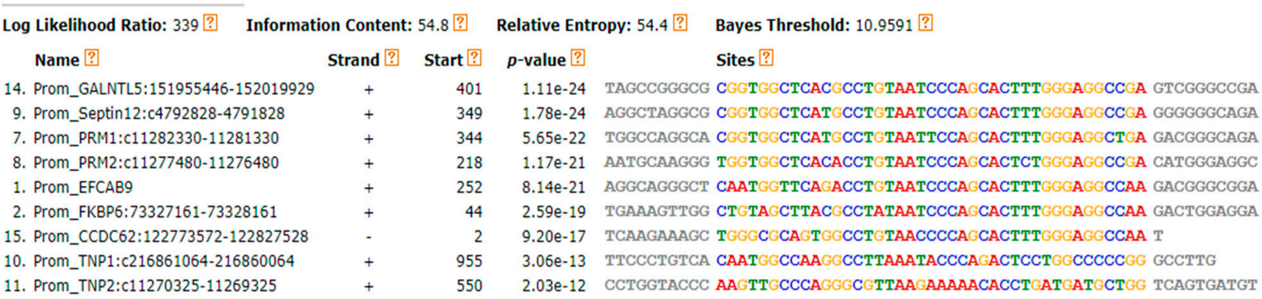


Figure 2. MEME Suite output showing sequence logos for the identified common promoter motif (Motif1) of genes predisposing nonsyndromic male infertility.

In 6 (six) genes (GALNTL5, SEPTIN12, PRM1, PRM2, EFCAB9, and TNP2), Motif1 is located between positions: −800 and −400 from the TSS. In 2 (two) genes, the motif is even further located: FKBP6 (−956) and CCDC62 (−998). CCDC62 is the only gene where the Motif1 site was found in the reverse complement of the supplied sequence. In the TNP1 gene, only 5 (five) nucleotides separate Motif1's end and the TSS of that gene, Figure 3.



Log Likelihood Ratio: 339 ? **Information Content:** 54.8 ? **Relative Entropy:** 54.4 ? **Bayes Threshold:** 10.9591 ?

	Name ?	Strand ?	Start ?	*p*-value ?	Sites ?
14.	Prom_GALNTL5:151955446-152019929	+	401	1.11e-24	TAGCCGGGCG CGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGA GTCGGGCCGA
9.	Prom_Septin12:c4792828-4791828	+	349	1.78e-24	AGGCTAGGCG CGGTGGCTCATGCCTGTAATCCCAGCACTTTGGGAGGCCGA GGGGGGCAGA
7.	Prom_PRM1:c11282330-11281330	+	344	5.65e-22	TGGCCAGGCA CGGTGGCTCATGCCTGTAATTCCAGCACTTTGGGAGGCTGA GACGGGCAGA
8.	Prom_PRM2:c11277480-11276480	+	218	1.17e-21	AATGCAAGGG TGGTGGCTCACACCTGTAATCCCAGCACTCTGGGAGGCCGA CATGGGAGGC
1.	Prom_EFCAB9	+	252	8.14e-21	AGGCAGGGCT CAATGGTTCAGACCTGTAATCCCAGCACTTTGGGAGGCCAA GACGGGCGGA
2.	Prom_FKBP6:73327161-73328161	+	44	2.59e-19	TGAAAGTTGG CTGTAGCTTACGCCTATAATCCCAGCACTTTGGGAGGCCAA GACTGGAGGA
15.	Prom_CCDC62:122773572-122827528	-	2	9.20e-17	TCAAGAAAGC TGGGCGCAGTGGCCTGTAACCCCAGCACTTTGGGAGGCCAA T
10.	Prom_TNP1:c216861064-216860064	+	955	3.06e-13	TTCCCTGTCA CAATGGCCAAGGCCTTAAATACCCAGACTCCTGGCCCCCGG GCCTTG
11.	Prom_TNP2:c11270325-11269325	+	550	2.03e-12	CCTGGTACCC AAGTTGCCCAGGGCGTTAAGAAAAACACCTGATGATGCTGG TCAGTGATGT

Figure 3. Motif1 sequence site with the 10 (ten) flanking letters on either side and the position in the sequence where the motif site starts.

Although Motif5, "GGCAGGAGRAKGGCBTGARCCCDGGRGGCRGMGSYTGCWGT" (Figure 4), is the motif with lowest predictive score [37], it is interesting to observe. Letters R, K, B, D, M, S, Y, W stand for an equal distribution on A/G, T/G, G/C/T, G/A/T, A/C, C/G, T/G, and A/T content, respectively. Nucleotides with a constant presence in all 9 (nine) genes that contain the motif are A on position 7, T on position 16, and G on position 31. Motif1 and Motif5 are the most frequent promoter common sequences, with a top hit rate of 64,2% in 17 (seventeen) examined genes associated with nonsyndromic male infertility. Motif5 was mainly distributed between positions −700 and −400 bp from the TSS (EFCAB9, SEPTIN12, GALANT5, PRM1, FKBP6, PRM2), and it was found almost proximal to the TSS of 3 (three) genes: HSF2(−155), TNP1(−194), and KLHL10(−78). Our findings clearly indicate that Motif5 comes in close proximity to the TSS only in 3 (three) genes, HSF2, TNP1, and KLHL10, while being distributed away from position −400 in the remaining genes, Figure 5.

The most reliable prediction for a common motif was Motif1, which serves as the most-likely binding site for transcription factors involved in gene regulation and expression.

We have performed further analysis in order to gain deeper insights into Motif1's pattern. This pattern was then compared against motifs cataloged in publicly accessible databases, to determine potential similarities with known regulatory motifs for transcription factors (TFs), using the TOMTOM web application [41]. Accordingly, Motif1 matched with 9 (nine) known motifs documented in databases, Table 3.

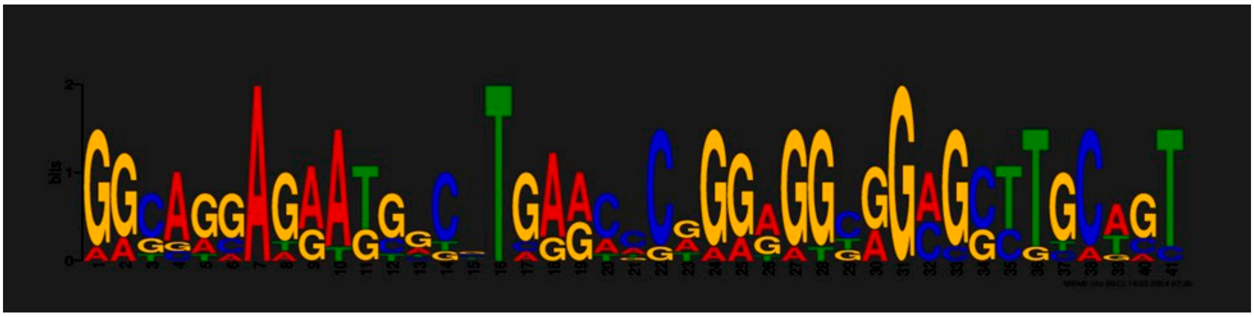
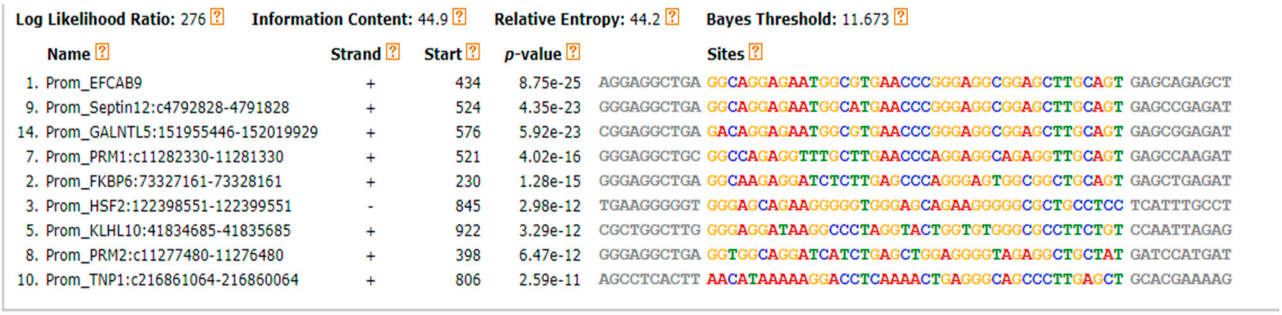


Figure 4. MEME Suite output showing sequence logos for the identified common motif (Motif5) of genes predisposing nonsyndromic male infertility.



```
Log Likelihood Ratio: 276 ?   Information Content: 44.9 ?   Relative Entropy: 44.2 ?   Bayes Threshold: 11.673 ?

   Name ?                                Strand ?  Start ?  p-value ?          Sites ?
 1. Prom_EFCAB9                             +        434    8.75e-25  AGGAGGCTGA GGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGT GAGCAGAGCT
 9. Prom_Septin12:c4792828-4791828          +        524    4.35e-23  GGGAGGCTGA GGCAGGAGAATGGCATGAACCCGGGAGGCGGAGCTTGCAGT GAGCCGAGAT
14. Prom_GALNTL5:151955446-152019929        +        576    5.92e-23  CGGAGGCTGA GACAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGT GAGCGGAGAT
 7. Prom_PRM1:c11282330-11281330            +        521    4.02e-16  GGGAGGCTGC GGCCAGAGGTTTGCTTGAACCCAGGAGGCAGAGGTTGCAGT GAGCCAAGAT
 2. Prom_FKBP6:73327161-73328161            +        230    1.28e-15  GGGAGGCTGA GGCAAGAGGATCTCTTGAGCCCAGGGAGTGGCGGCTGCAGT GAGCTGAGAT
 3. Prom_HSF2:122398551-122399551           -        845    2.98e-12  TGAAGGGGGT GGGAGCAGAAGGGGGTGGGAGCAGAAGGGGGCGCTGCCTCC TCATTTGCCT
 5. Prom_KLHL10:41834685-41835685           +        922    3.29e-12  CGCTGGCTTG GGGAGGATAAGGCCCTAGGTACTGGTGTGGGCGCCTTCTGT CCAATTAGAG
 8. Prom_PRM2:c11277480-11276480            +        398    6.47e-12  GGGAGGCTGA GGTGGCAGGATCATCTGAGCTGGAGGGGTAGAGGCTGCTAT GATCCATGAT
10. Prom_TNP1:c216861064-216860064          +        806    2.59e-11  AGCCTCACTT AACATAAAAAGGACCTCAAAACTGAGGGCAGCCCTTGAGCT GCACGAAAAG
```

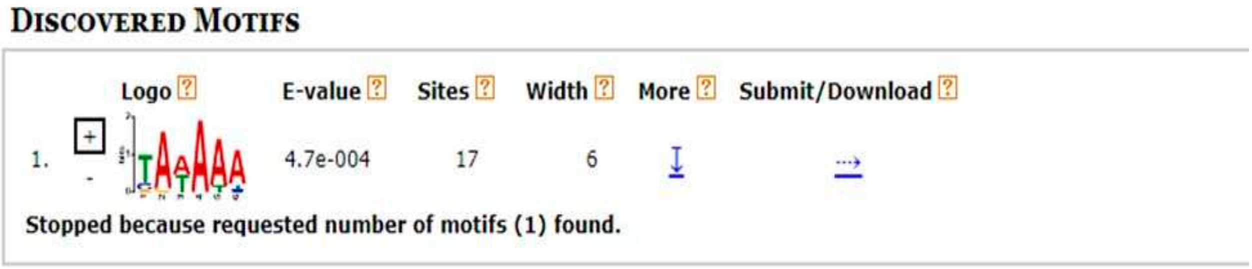
Figure 5. Motif5 sequence site with the 10 (ten) flanking letters on either side and the position in the sequence where the motif site starts.

Table 3. List of matching transcription factors (TFs) which could bind Motif1.

Gene ID	Gene Name	Species	TF Family	Candidate TF	Statistical Significance
GLIS1	GLIS1.H12CORE.0.P.B	*Homo sapiens*	More than 3 adjacent zinc finger factors	C2H2 zinc finger factors	5.43e-04
ZSCAN21	ZSC21.H12CORE.0.P.C	*Homo sapiens*	More than 3 adjacent zinc finger factors	C2H2 zinc finger factors	6.58e-04
GLIS3	GLIS3.H12CORE.0.P.C	*Homo sapiens*	More than 3 adjacent zinc finger factors	C2H2 zinc finger factors	9.23e-04
GLIS1	GLIS1.H12CORE.1.P.B	*Homo sapiens*	More than 3 adjacent zinc finger factors	C2H2 zinc finger factors	2.67e-03
ZNF770	ZN770.H12CORE.0.P.B	*Homo sapiens*	Multiple dispersed zinc fingers	C2H2 zinc finger factors	2.92e-03
ZNF780A	Z780A.H12CORE.0.P.C	*Homo sapiens*	More than 3 adjacent zinc finger factors	C2H2 zinc finger factors	3.52e-03
ZNF81	ZNF81.H12CORE.0.P.C	*Homo sapiens*	More than 3 adjacent zinc finger factors	C2H2 zinc finger factors	4.51e-03
ZNF264	ZN264.H12CORE.0.P.B	*Homo sapiens*	More than 3 adjacent zinc finger factors	C2H2 zinc finger factors	4.98e-03
JUNB	JUNB.H12CORE.0.PM.A	*Homo sapiens*	Jun-related	Basic leucine zipper factors (bZIPs)	5.08e-03

We have found that Motif1 bears a significant resemblance to the binding motif recognized by zinc finger (ZNF) transcription factors in 8 (eight) genes, GLIS1, ZSCAN21, GLIS3, GLIS1, ZNF770, ZNF780A, ZNF81, and ZNF264, suggesting the common regulation mechanism of these genes. On the other hand, in 1 (one) gene, JUNB, Motif1 serves as a binding site for basic leucine zipper factors (bZIPs), indicating a unique regulation property among the genes associated with nonsyndromic male infertility. Hence, it is plausible that Motif1 could function as a binding site for ZNF and bZIP TFs in humans, thereby regulating the expression of these genes.

We also applied an alternative approach to identify a promoter consensus sequence shared among the target genes. The predicted promoters with the highest NNPP score were collected for each of the 17 (seventeen) genes. These sequences were then analyzed using the MEME program. Applying MEME version 5.5.5, the application successfully identified 1 (one) statistically significant consensus sequence, TAWAAA (E-value: 4.7e-004), Figure 6, which was present in all 17 (seventeen) gene promoters of interest. Promoter consensus sequence: TAWAAA, such as T stands for Thymine, A for Adenine, and W for equal A/T appearance, has a width of 6 (six) nucleotides and appeared at 17 (seventeen) distinct sites. It was observed on the positive strand in 15 (fifteen) sequences and on the negative strand in 2 (two) sequences. Across the majority of sequences, TAWAAA or the TATA box is positioned between positions 11 and 17 within the 50 bp promoter regions of the highest scores, except for gene EFCAB9, where it starts at position 1, Figure 7. This finding implicates a potentially different regulation mechanism of the gene EFCAB9.

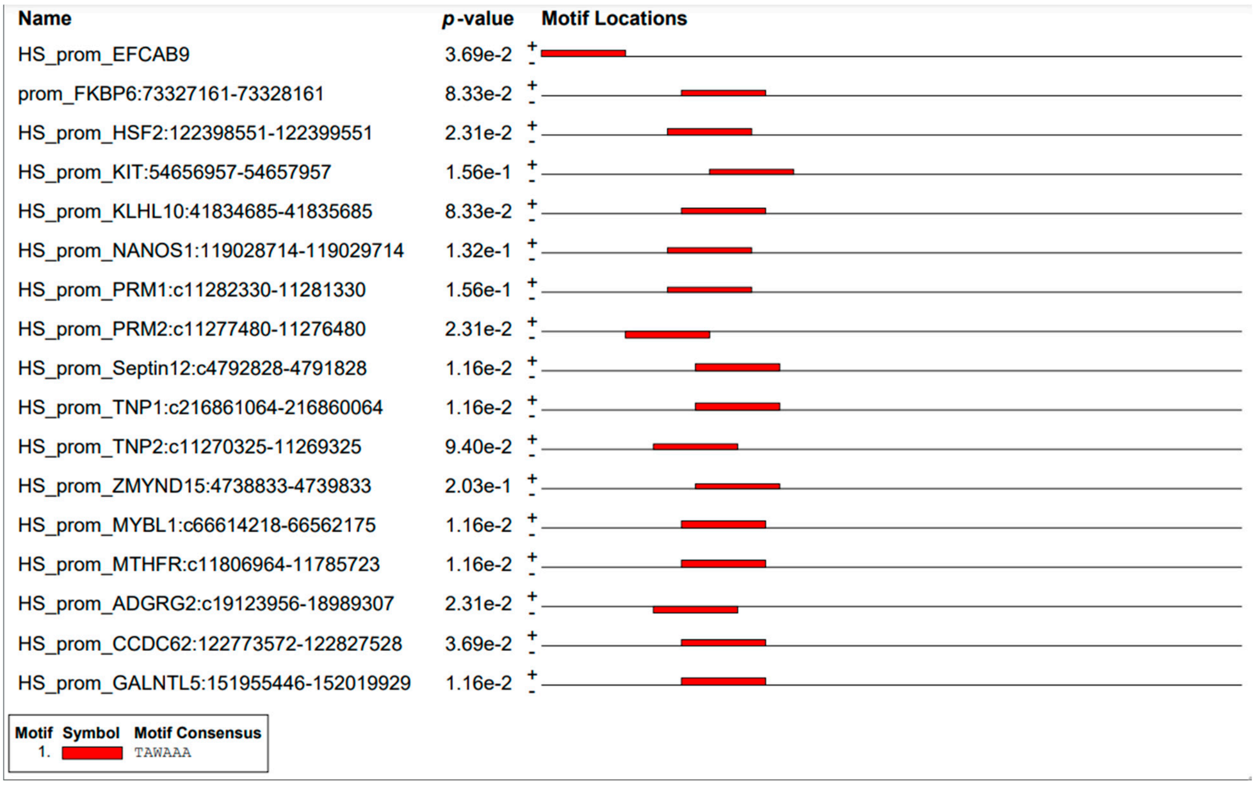


DISCOVERED MOTIFS

	Logo	E-value	Sites	Width	More	Submit/Download
1.	+ - TAWAAA	4.7e-004	17	6	↧	⇢

Stopped because requested number of motifs (1) found.

Figure 6. MEME Suite output—consensus sequence "TAWAAA" (TATA box) found among the highest scoring promoter regions.



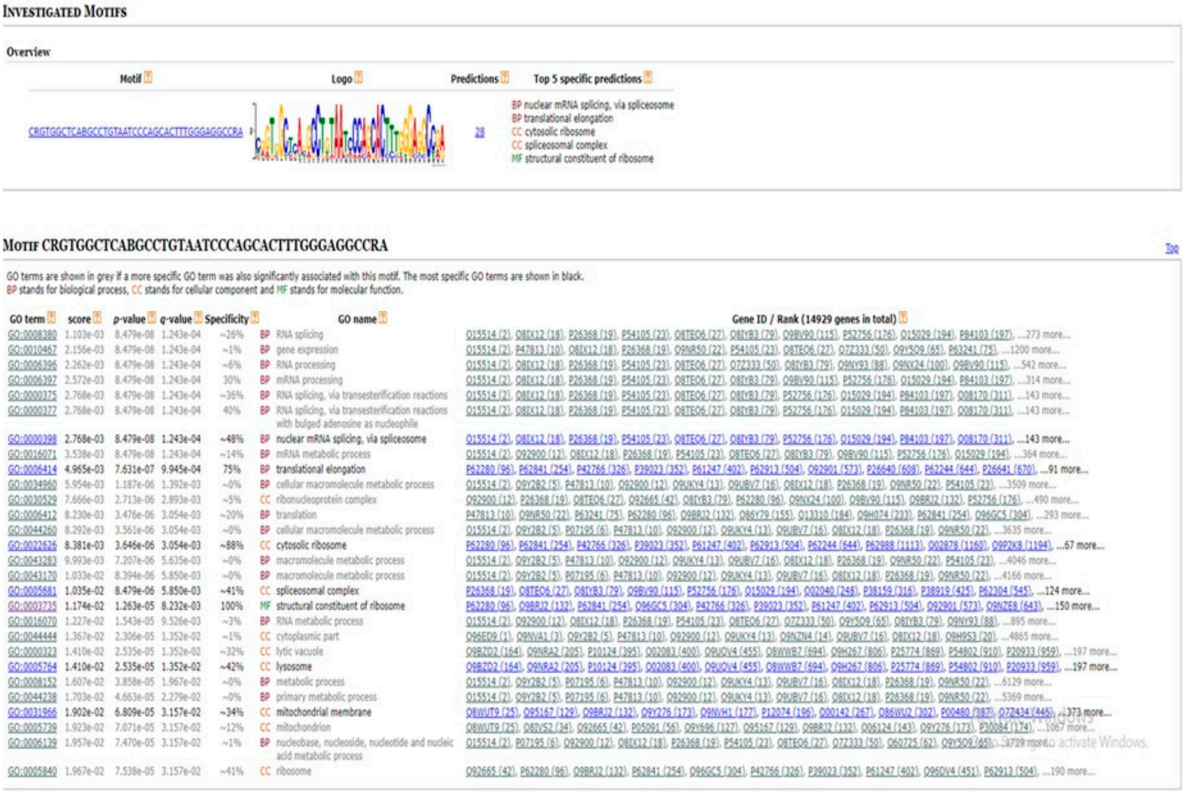
Name	*p*-value	Motif Locations
HS_prom_EFCAB9	3.69e-2	+ -
prom_FKBP6:73327161-73328161	8.33e-2	+ -
HS_prom_HSF2:122398551-122399551	2.31e-2	+ -
HS_prom_KIT:54656957-54657957	1.56e-1	+ -
HS_prom_KLHL10:41834685-41835685	8.33e-2	+ -
HS_prom_NANOS1:119028714-119029714	1.32e-1	+ -
HS_prom_PRM1:c11282330-11281330	1.56e-1	+ -
HS_prom_PRM2:c11277480-11276480	2.31e-2	+ -
HS_prom_Septin12:c4792828-4791828	1.16e-2	+ -
HS_prom_TNP1:c216861064-216860064	1.16e-2	+ -
HS_prom_TNP2:c11270325-11269325	9.40e-2	+ -
HS_prom_ZMYND15:4738833-4739833	2.03e-1	+ -
HS_prom_MYBL1:c66614218-66562175	1.16e-2	+ -
HS_prom_MTHFR:c11806964-11785723	1.16e-2	+ -
HS_prom_ADGRG2:c19123956-18989307	2.31e-2	+ -
HS_prom_CCDC62:122773572-122827528	3.69e-2	+ -
HS_prom_GALNTL5:151955446-152019929	1.16e-2	+ -

Motif	Symbol	Motif Consensus
1.	🟥	TAWAAA

Figure 7. Highest scoring promoter regions and TATA box distribution within.

3.3. Gene Ontology for MOTIF1

We tried to identify Gene Ontology (GO) terms for the Motif1 common promoter sequence. This was accomplished using the GOMo (Gene Ontology for Motifs) application, version 5.5.5 [42]. We have found in total 28 (twenty-eight) GO predictions with different functions, Figure 8. The most specific GO terms associated with Motif1 are biological processes, such as nuclear mRNA splicing via spliceosome (48% specificity) and translational elongation (75%); cellular components, including the cytosolic ribosome (88%), spliceosomal complex (41%), lysosome (42%), and mitochondrial membrane (34%); and molecular function, including the structural constituent of the ribosome, Figure 8. For the GO term with 100% specificity in prediction, GO:0003735 has a molecular function and is found as a structural constituent of the ribosome; prediction score = 1.174e-02, p-value = 1.068e-05 and q-value = 6.962e-03. It is defined as the action of a molecule that contributes to the structural integrity of the ribosome and is related to all genes and gene products annotated to structural constituent of ribosome and all direct and indirect annotations to structural constituent of ribosome. It has a total 33983 annotations in papers in the Eukaryota taxonomic group, where In Homo Sapiens has 6069 annotations: 5269 as rRNA (16S mitochondrial, 5S, 5.8S, 12S, 16S, 18S and 28S ribosomal rRNA), 415 as protein (large and small ribosomal subunit protein, NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, putative ribosomal protein), 377 as gene product (coding and non-protein coding), 5 as ncRNA (large and small ribosomal subunit), 2 as tRNA (HSALNT0258931 and spliced polyadenylated non-coding RNA), and 1 as snoRNA (partial 5SN1 small nucleolar RNA).



INVESTIGATED MOTIFS

Overview

Motif	Logo	Predictions	Top 5 specific predictions
CRGTGGCTCABGCCTGTAATCCCAGCACTTTGGGAGGCCRA		28	BP nuclear mRNA splicing, via spliceosome BP translational elongation CC cytosolic ribosome CC spliceosomal complex MF structural constituent of ribosome

MOTIF CRGTGGCTCABGCCTGTAATCCCAGCACTTTGGGAGGCCRA

Top

GO terms are shown in grey if a more specific GO term was also significantly associated with this motif. The most specific GO terms are shown in black.
BP stands for biological process, CC stands for cellular component and MF stands for molecular function.

GO term	score	p-value	q-value	Specificity		GO name	Gene ID / Rank (14929 genes in total)
GO:0008380	1.103e-03	8.479e-08	1.243e-04	~26%	BP	RNA splicing	O15514 (2), Q8IX12 (18), P26368 (19), P54105 (23), Q8TEQ6 (27), Q8IYB3 (79), Q9BV90 (115), P52756 (176), Q15029 (194), P84103 (197), ...273 more...
GO:0010467	2.156e-03	8.479e-08	1.243e-04	~1%	BP	gene expression	O15514 (2), P47813 (10), Q8IX12 (18), P26368 (19), Q9NR50 (22), P54105 (23), Q8TEQ6 (27), Q7Z333 (50), Q9Y5Q9 (65), P63241 (75), ...1200 more...
GO:0006396	2.262e-03	8.479e-08	1.243e-04	~6%	BP	RNA processing	O15514 (2), Q8IX12 (18), P26368 (19), P54105 (23), Q8TEQ6 (27), Q7Z333 (50), Q8IYB3 (79), Q9NY93 (88), Q9NX24 (100), Q9BV90 (115), ...542 more...
GO:0006397	2.572e-03	8.479e-08	1.243e-04	30%	BP	mRNA processing	O15514 (2), Q8IX12 (18), P26368 (19), P54105 (23), Q8TEQ6 (27), Q8IYB3 (79), Q9BV90 (115), P52756 (176), Q15029 (194), P84103 (197), ...314 more...
GO:0000375	2.768e-03	8.479e-08	1.243e-04	~36%	BP	RNA splicing, via transesterification reactions	O15514 (2), Q8IX12 (18), P26368 (19), P54105 (23), Q8TEQ6 (27), Q8IYB3 (79), P52756 (176), Q15029 (194), P84103 (197), Q08170 (311), ...143 more...
GO:0000377	2.768e-03	8.479e-08	1.243e-04	40%	BP	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	O15514 (2), Q8IX12 (18), P26368 (19), P54105 (23), Q8TEQ6 (27), Q8IYB3 (79), P52756 (176), Q15029 (194), P84103 (197), Q08170 (311), ...143 more...
GO:0000398	2.768e-03	8.479e-08	1.243e-04	~48%	BP	nuclear mRNA splicing, via spliceosome	O15514 (2), Q8IX12 (18), P26368 (19), P54105 (23), Q8TEQ6 (27), Q8IYB3 (79), P52756 (176), Q15029 (194), P84103 (197), Q08170 (311), ...143 more...
GO:0016071	3.538e-03	8.479e-08	1.243e-04	~14%	BP	mRNA metabolic process	O15514 (2), Q92900 (12), Q8IX12 (18), P26368 (19), P54105 (23), Q8TEQ6 (27), Q8IYB3 (79), Q9BV90 (115), P52756 (176), Q15029 (194), ...364 more...
GO:0006414	4.965e-03	7.631e-07	9.945e-04	75%	BP	translational elongation	P62280 (96), P62841 (254), P42766 (326), P39023 (352), P61247 (402), P62913 (504), Q92901 (573), P26640 (608), P62244 (644), P26641 (670), ...91 more...
GO:0034960	5.954e-03	1.187e-06	1.392e-03	~0%	BP	cellular macromolecule metabolic process	O15514 (2), Q9Y2B2 (5), P47813 (10), Q92900 (12), Q9UKY4 (13), Q9UBV7 (16), Q8IX12 (18), P26368 (19), Q9NR50 (22), P54105 (23), ...3509 more...
GO:0030529	7.666e-03	2.713e-06	2.893e-03	~5%	CC	ribonucleoprotein complex	Q92900 (12), P26368 (19), Q8TEQ6 (27), Q92665 (42), Q8IYB3 (79), P62280 (96), Q9NX24 (100), Q9BV90 (115), Q9BRJ2 (132), P52756 (176), ...490 more...
GO:0006412	8.230e-03	3.476e-06	3.054e-03	~20%	BP	translation	P47813 (10), Q9NR50 (22), P63241 (75), P62280 (96), Q9BRJ2 (132), Q86Y79 (155), Q13310 (184), Q9H074 (233), P62841 (254), Q96GC5 (304), ...293 more...
GO:0044260	8.292e-03	3.561e-06	3.054e-03	~0%	BP	cellular macromolecule metabolic process	O15514 (2), Q9Y2B2 (5), P07195 (6), P47813 (10), Q92900 (12), Q9UKY4 (13), Q9UBV7 (16), Q8IX12 (18), P26368 (19), Q9NR50 (22), ...3635 more...
GO:0022626	8.381e-03	3.646e-06	3.054e-03	~88%	CC	cytosolic ribosome	P62280 (96), P62841 (254), P42766 (326), P39023 (352), P61247 (402), P62913 (504), P62244 (644), P62988 (1113), Q02878 (1160), Q9P2K8 (1194), ...67 more...
GO:0043283	9.993e-03	7.207e-06	5.635e-03	~0%	BP	macromolecule metabolic process	O15514 (2), Q9Y2B2 (5), P47813 (10), Q92900 (12), Q9UKY4 (13), Q9UBV7 (16), Q8IX12 (18), P26368 (19), Q9NR50 (22), P54105 (23), ...4046 more...
GO:0043170	1.033e-02	8.394e-06	5.850e-03	~0%	BP	macromolecule metabolic process	O15514 (2), Q9Y2B2 (5), P07195 (6), P47813 (10), Q92900 (12), Q9UKY4 (13), Q9UBV7 (16), Q8IX12 (18), P26368 (19), Q9NR50 (22), ...4166 more...
GO:0005681	1.035e-02	8.479e-06	5.850e-03	~41%	CC	spliceosomal complex	P26368 (19), Q8TEQ6 (27), Q8IYB3 (79), Q9BV90 (115), P52756 (176), Q15029 (194), Q02040 (248), P38159 (316), P38919 (425), P62304 (545), ...124 more...
GO:0003735	1.174e-02	1.263e-05	8.232e-03	100%	MF	structural constituent of ribosome	P62280 (96), Q9BRJ2 (132), P62841 (254), Q96GC5 (304), P42766 (326), P39023 (352), P61247 (402), P62913 (504), Q92901 (573), Q9NZE8 (643), ...150 more...
GO:0016070	1.227e-02	1.543e-05	9.526e-03	~3%	BP	RNA metabolic process	O15514 (2), Q92900 (12), Q8IX12 (18), P26368 (19), P54105 (23), Q8TEQ6 (27), Q7Z333 (50), Q9Y5Q9 (65), Q8IYB3 (79), Q9NY93 (88), ...895 more...
GO:0044444	1.367e-02	2.306e-05	1.352e-02	~1%	CC	cytoplasmic part	Q96ED9 (1), Q9NVA1 (3), Q9Y2B2 (5), P47813 (10), Q92900 (12), Q9UKY4 (13), Q9NZN4 (14), Q9UBV7 (16), Q8IX12 (18), Q9H9S3 (20), ...4865 more...
GO:0000323	1.410e-02	2.535e-05	1.352e-02	~32%	CC	lytic vacuole	Q9BZD2 (164), Q9NRA2 (205), P10124 (395), Q02083 (400), Q9UQV4 (455), Q8WWB7 (694), Q9H267 (806), P25774 (869), P54802 (910), P20933 (959), ...197 more...
GO:0005764	1.410e-02	2.535e-05	1.352e-02	~42%	CC	lysosome	Q9BZD2 (164), Q9NRA2 (205), P10124 (395), Q02083 (400), Q9UQV4 (455), Q8WWB7 (694), Q9H267 (806), P25774 (869), P54802 (910), P20933 (959), ...197 more...
GO:0008152	1.607e-02	3.858e-05	1.967e-02	~0%	BP	metabolic process	O15514 (2), Q9Y2B2 (5), P07195 (6), P47813 (10), Q92900 (12), Q9UKY4 (13), Q9UBV7 (16), Q8IX12 (18), P26368 (19), Q9NR50 (22), ...6129 more...
GO:0044238	1.703e-02	4.663e-05	2.279e-02	~0%	BP	primary metabolic process	O15514 (2), Q9Y2B2 (5), P07195 (6), P47813 (10), Q92900 (12), Q9UKY4 (13), Q9UBV7 (16), Q8IX12 (18), P26368 (19), Q9NR50 (22), ...5369 more...
GO:0031966	1.902e-02	6.809e-05	3.157e-02	~34%	CC	mitochondrial membrane	Q8WUT9 (25), O95167 (129), Q9BRJ2 (132), Q9Y276 (173), Q9NVH1 (177), P12074 (196), O00142 (267), Q86WU2 (302), P00480 (387), O72434 (445), ...373 more...
GO:0005739	1.923e-02	7.071e-05	3.157e-02	~12%	CC	mitochondrion	Q8WUT9 (25), Q8IVS2 (34), Q92665 (42), P05091 (56), Q9Y696 (127), O95167 (129), Q9BRJ2 (132), Q06124 (143), Q9Y276 (173), P30084 (174), ...1067 more...
GO:0006139	1.957e-02	7.470e-05	3.157e-02	~1%	BP	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	O15514 (2), P07195 (6), Q92900 (12), Q8IX12 (18), P26368 (19), P54105 (23), Q8TEQ6 (27), Q7Z333 (50), O60725 (62), Q9Y5Q9 (65), ...1729 more...
GO:0005840	1.967e-02	7.538e-05	3.157e-02	~41%	CC	ribosome	Q92665 (42), P62280 (96), Q9BRJ2 (132), P62841 (254), Q96GC5 (304), P42766 (326), P39023 (352), P61247 (402), Q96DV4 (451), P62913 (504), ...190 more...

Figure 8. List of Gene Ontology (GO) terms specific to Motif 1.

3.4. CpG Islands in Promoters and Gene Body Regions of Genes Associated with Nonsyndromic Male Infertility

CpG islands (CGIs) typically occur towards the 5′ end of genes and contain dinucleotides rich in GC content. There is a noted association between a lower frequency of methylation and higher CpG density, and contrariwise [44]. DNA methylation involves substituting the hydrogen attached to a cytosine base with a methyl group, leading to an increased chromatin compaction that impacts the binding of transcription factors [45]. If "gene body" is defined as the entire gene from the transcription start site (TSS) to the end of the transcript [46], the sequence 1 kb upstream from the transcription start site (TSS) we define as the "promoter region". Genome-wide methylation studies have shown that DNA methylation is widespread not only in promoters but also in gene bodies [47]. We searched for CpG islands in both the "promoter region" and "gene body" using the database of CpG islands and the analytical tool (DBCAT) "http://dbcat.cgm.ntu.edu.tw (accessed on 1 March 2024)", with search criteria of GC content $\geq$ 55%, Observed CpG/Expected CpG ratio $\geq$ 0.65, and length $\geq$ 500 bp. Accordingly, for the analysis of the "promoter region" and "gene body" segments of all 17 genes, the program did not find any CpG island in 5 genes: EFCAB9, PRM1, PRM2, TNP1, and TNP2. The majority of the 12 genes have one (1) CpG island located in the gene body (43.7%) and the rest have from 2 to 10 CpG islands, such as in the SEPTIN12 and KLHL10 genes, Table 4. The length of the CpG islands in the gene body is from 521 to 1616 bp, which is about 50–68% of the CG content, Table 4. The CpG islands are usually located at the beginning of the gene body, close to the TSS, except in the GALNTL5 gene, where the first CpG Island starts at the 38 166 position downstream of the TSS. When searching for CpG islands in the regions 1 kb upstream from the TSS of the genes, the algorithm did not find any in 5 (five) genes, HSF2, KLHL10, CCDC62, ADGRG2, and GALNTL5, while all the rest have one CpG island located in the promoter regions, Table 4. The majority of the CpG islands are located at the endings of the promoter regions, close to the TSS, except in SEPTIN12, where the CpG island is located at the beginning of the promoter region. All 7 (seven) promoter regions are rich in GC content (56–69%) and in some cases with a length of 1 kb (Table 4).

Table 4. Number of CpG islands identified and fragment sizes for genes predisposing nonsyndromic male infertility.

Region	Sequence Name	No. CpG Islands Discovered (Start Location)	Fragment Sizes (>500 bp)	% GC Content
1 kb upstream from the TSS	Prom_FKBP6	1 (331)	720	62%
	Prom_KIT	1 (731)	1009	61%
	Prom_NANOS1	1 (119)	947	69%
	Prom_ZMYND15	1 (86)	997	67%
	Prom_MYBL1	1 (77)	1007	59%
	Prom_SEPT12	1 (99)	590	56%
	Prom_MTHFR	1 (399)	685	64%
Gene body	FKBP6	2 (82, 26751)	855, 640	58%, 51%
	HSF2	1 (82)	1006	59%
	KIT	1 (821)	1666	62%
	KLHL10	1 (961)	1162	65%
	NANOS1	1 (77)	1584	65%
	SEPT12	3 (5151, 7593, 12776)	759, 869, 521	53%, 63%, 50%
	ZMYND15	2 (75, 5119)	655, 1172	60%, 65%
	MYBL1	1 (90)	1105	59%
	CCDC62	4 (80, 8559, 40725, 50782)	911, 521, 527, 994	56%, 50%, 51%, 50%
	ADGRG2	2 (81, 71660)	1616, 565	68%, 53%
	MTHFR	1 (78)	1011	61%
	GALNTL5	1 (38166)	865	51%

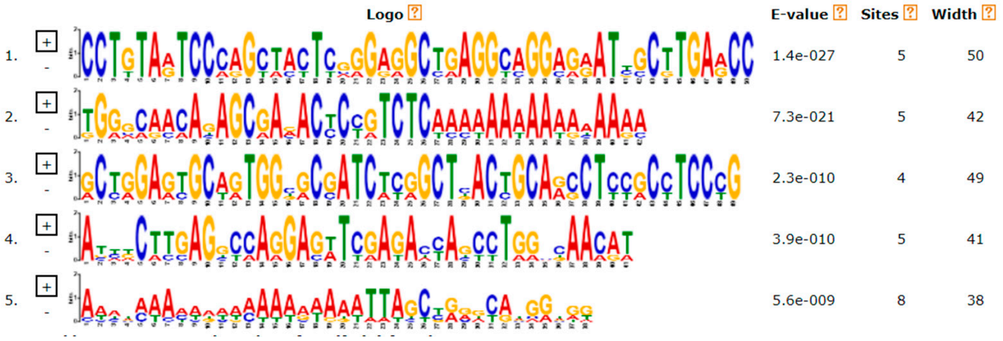
3.5. Validation

To verify the reliability of our findings, we have selected 10 (ten) human housekeeping genes, GAPDH, PGK1, PPIA, RPL13A, RPLP0, B2M, SDHA, GUSB, HMBS, and TBP, which are not associated with infertility, as a negative control group. These sequences were retrieved in FASTA format from the National Center for Biotechnology Information (NCBI) Genome Browser "https://www.ncbi.nlm.nih.gov/gene (accessed on 1 March 2024)" and analyzed, applying the same methodology and applications as we did for our target genes. We used the NNPP application to search for genes' promoters within the region of 1 kb upstream of the genes' known transcription start site (TSS) (Table 5), at a promoter predictivity cut off value of 0.8. The number of identified promoters ranged between 1 and 4, Table 5. We used the MEME application to search for common promoter motifs and the top 5 (five) hits were reported, further referred to as hkg_cpm1, hkg_cpm2, hkg_cpm3, hkg_cpm4, and hkg_cpm5 (Figure 9).

Table 5. Predictive score and number of promoters for 10 (ten) human housekeeping genes.

Gene Symbol (Gene ID) Full Name *	Corresponding Promoter Region Name	No. of Promoters Identified in Promoter Region (1000 bp Upstream)	Predictive Score at Cut Off Value 0.8 **
GAPDH (Gene ID: 2597) glyceraldehyde-3-phosphate dehydrogenase	Prom_GAPDH	4	0.84, 1.0, 0.87, 1.0
PGK1 (ID: 5230) phosphoglycerate kinase 1	Prom_PGK1	3	0.98, 0.97, 1.0
PPIA (ID: 5478) Peptidylprolyl isomerase A	Prom_PPIA	2	0.91, 1.0
RPL13A (ID: 23521) ribosomal protein L13a	Prom_RPL13A	3	0.98, 1.0, 1.0
RPLP0 (ID:285588) Ribosomal protein, large, P0	Prom_RPLP0	3	0.89, 1.0, 0.88
B2M (ID: 567) Beta-2-microglobulin	Prom_B2M	1	1.0
SDHA (ID: 6389) Succinate dehydrogenase complex, subunit A, flavoprotein (Fp)	Prom_SDHA	3	0.95, 0.99, 0.99
GUSB (ID: 2990) glucuronidase beta	Prom_GUSB	3	0.86, 0.89, 0.82
HMBS (ID: 3145) Hydroxymethylbilane synthase	Prom_HMBS	2	0.96, 0.82
TBP (ID: 6908) TATA box binding protein	Prom_TBP	1	0.95

* provided by HGNC "https://www.genenames.org/ (accessed on 1 March 2024)"; ** Cut off value is set to 0.8 for reliable predictions.



	Logo	E-value	Sites	Width
1.		1.4e-027	5	50
2.		7.3e-021	5	42
3.		2.3e-010	4	49
4.		3.9e-010	5	41
5.		5.6e-009	8	38

Figure 9. Human housekeeping genes' common promoter motifs: hkg_cpm [1–5], identified by MEME application.

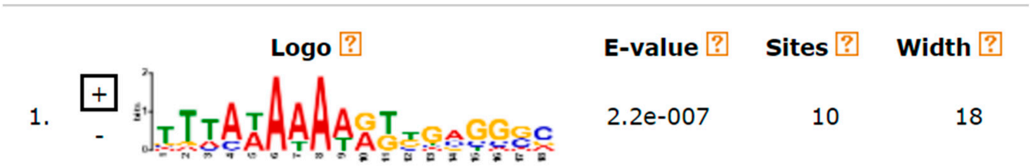
We used EMBOSS Needle aligner [48] to compare the common promoter Motif1="CRGTGGCTCABGCCTGTAATCCCAGCACTTTGGGAGGCCRA", which is specific to genes associated with nonsyndromic male infertility, to all 5 (five) human housekeeping genes' common promoter motifs shown on Figure 9. Table 6 reports on the percentage of similarity between Motif 1 and hkg_cpm [1–5], employing a gap opening penalty of 10 and gap extension penalty of 0.5. The highest similarity percentage of 41.3% was obtained for hkg_cpm1="CCTGTARTCCCAGCTACTCGGGAGGCTGAGGCAGGAGRATBGCTTGARCC", while the lowest was for hkg_cpm5, Table 6.

Table 6. Similarity percentage (%) between Motif 1 and hkg_cpm [1–5].

Common Promoter Motif	Motif 1
hkg_cpm1	41.3%
hkg_cpm2	12.5%
hkg_cpm3	29.9%
hkg_cpm4	28.8%
hkg_cpm5	6.8%

The obtained results clearly show that Motif 1 and hkg_cpm [1–5] are highly dissimilar, unrelated sequences, validating Motif 1 as a common promoter motif which is specific to nonsyndromic male infertility-associated genes exclusively.

We have also identified a consensus motif for the 10 (ten) randomly taken human housekeeping genes (GAPDH, PGK1, PPIA, RPL13A, RPLP0, B2M, SDHA, GUSB, HMBS, TBP), unrelated to infertility. The consensus motif, TTTAWAAAARKBGMGGSC (Figure 10), with a length of 18 (eighteen) base pairs and being present in all 10 (ten) negative control genes, was identified. Our findings prove that the common form of the TATA box, which is TAWAAA, such as W is A or T, can be exclusively attributed to the group of nonsyndromic male infertility-associated genes.



		Logo ?	E-value ?	Sites ?	Width ?
1.	+ -		2.2e-007	10	18

Figure 10. Human housekeeping genes' consensus motif.

4. Discussion

Other studies also analyze the genes of interest in this article. For instance, Guerri et al. [30] analyzed new candidate genes that might be responsible for male infertility resulting from single-gene mutations, such that they developed an NGS panel to detect nucleotide variations in the coding exons and flanking regions of all the genes associated with infertility. Given that male infertility is suspected, Guerri and colleagues [30] analyze the same set of genes as we do. However, there is a conceptual distinction between [30] and our study. Guerri et al. [30] analyze mutagenesis contributing to nonsyndromic male infertility, while we analyze the properties of the expression mechanisms of genes linked to nonsyndromic male infertility. Guerri and colleagues [30] reported on pathogenic missense, nonsense, and splicing mutations that cause azoospermia, macrozoospermia, globozoospermia, and other conditions of sperm defects and nonsyndromic male infertility. Another study [49] reviews the most common autosomal recessive and autosomal dominant single-gene disorders involved in human infertility. The genes covered inside are SPATA16, AURKC, CATSPER1, MTHFR, and SYCP3. Okutman et al. [31] emphasize the challenges of studying patient cohorts due to the multiple possible causes of male infertility, both genetic and non-genetic, and the limited discernment of diagnostic tests. Phenotypic homogeneity is a major paradigm in sporadic cases. Azoospermia has various causes, making it very difficult, if not impossible, to classify them. According to Okutman et al. [31], there are 17 (seventeen) human genes that, when mutated, lead to severe nonsyndromic oligozoospermia and/or azoospermia without overlapping with female infertility. The genes of interest in this study were also considered by Zorrilla and Yatsenko [50].

All these papers, and many more, study the impact of particular mutations upon the function of genes, given that nonsyndromic male infertility has been confirmed on an individual basis. On the other hand, our study aims to identify common and unique transcriptional properties of genes associated with nonsyndromic male infertility.

The transcription start site (TSS) refers to the first nucleotide being transcribed, while the nearby genomic region of the TSS is often referred to as the core promoter [51]. Upon receiving the right external signals, the core promoter takes part in the formation of a transcription preinitiation complex alongside various accessory proteins, such as RNA polymerase and transcription factors, helping the initiation of transcription [51–55]. The regulation of transcriptional initiation is a critical step in the control of gene expression [56,57]. The transcription of a gene may start from one of several TSSs, a phenomenon known as alternative transcriptional initiation (ATI), and the different core promoters used are alternative promoters [55,56]. It has been reported that ATI occurs to most eukaryotic protein-coding genes [56–62]. For example, over 50% of all human genes have alternative promoters [61], and on average, a human gene has 4 (four) TSSs [57]. ATI enables the generation of transcripts from the same gene that vary in their 5′ untranslated region (5′ UTR) or even the protein-coding region [63].

The identification of the transcription start site was a crucial point in addition to this study, as we aimed to advance the identification and characterization of promoter regions where significant regulatory elements are expected to bind, playing a pivotal role in gene regulation [64]. This analysis found that 65,4% of the genes associated with nonsyndromic male infertility have 1 (one) to 6 (six) promoters 1 kb upstream of the TSS. Dai and colleagues [65] suggested that genes attributed with multiple promoters increase the likelihood of transcription initiation and contribute to gene expression in response to changes in environmental conditions. This finding agrees with our findings. The location of the majority of identified promoters was at ≤ -500 bp from the start codon.

Transcriptional factors modulate gene expression through binding to a specific DNA sequence, usually found upstream of the gene, or the genomic region that they control [12]. There are several known transcription factors involved in the expression of nonsyndromic male infertility-associated genes, such as DMRT (Doublesex and Mab-3 Related Transcription factor), SOX9, HOXA10, FOXJ1, and Zinc Finger Proteins [29,32,66]. On the other hand, WT1 (Wilms tumor 1 protein), Steroidogenic factor-1 (SF-1), and FOXL2 (Forkhead box protein L2) were proved to be involved in the expression of syndromic male infertility-associated genes [67–71]. Our study reports C2H2 zinc finger protein as a common transcription factor of nonsyndromic male infertility-associated genes. However, there is an exception when it comes to the JUNB gene, which is transcribed by basic leucine zipper factors (bZIPs) specifically.

Motif1, “CRGTGGCTCABGCCTGTAATCCCAGCACTTTGGGAGGCCRA”, was identified as the most reliable common promoter motif for genes associated with nonsyndromic male infertility, which serves as a binding site for C2H2 zinc finger (ZNF) transcription factors, to regulate the expression of these genes. Although C2H2 zinc finger TF was found as a common and most significant transcription factor, binding promoters in several genes such as GLIS1, ZSCAN21, GLIS3, GLIS1, ZNF770, ZNF780A, ZNF81, and ZNF264, we have found that Motif1 also serves as a biding site for basic leucine zipper factors (bZIPs) in JUNB gene. Our finding suggests that the JUNB gene might have different regulation properties compared to the other genes, which remains to be experimentally verified in the future.

The most significant common motif, Motif1, was found to be associated with 28 (twenty-eight) Gene Ontology terms, including biological processes, such as nuclear mRNA splicing (via spliceosome) and translational elongation. Our study has also found that the highest-ranking promoter predictions share a common TATA box consensus sequence, TAWAAA, such as W is A or T. We have also found that the TATA box in the EFCAB9 gene is located at least 10 bp away from its common position in the rest of the genes associated with nonsyndromic male infertility.

The CpG analysis showed that in total 12 (twelve) genes associated with nonsyndromic infertility have at least 1 (one) CpG island in the “gene body” and 7 (seven) of them have a CpG island in the promoter region. The top 2 (two) genes, with the highest CpG density in their promoter regions and a fragment size of approximately 1 kb, are NANOS1 (fragment size = 947 bp, %CG content = 69%) and ZMYND15 (fragment size = 997 bp, %CG content = 67%), which are expected to be less susceptible to DNA methylation, compared to the other genes associated with nonsyndromic male infertility, reducing the malfunctioning risk.

There are also certain limitations, in addition, to our in silico analysis. Gene mutations, such as SNPs or indels, which usually interfere with TF binding activity, are not considered in this study, as we primarily aim to analyze the control mechanisms of the regulation of genes associated with nonsyndromic male infertility. There are also limitations in addition to the used in silico applications. The NNPP program (Neural Network Promoter Prediction) implements a time delay neural network, and the accuracy of the prediction depends on the amount of gene data used to train the model. Accordingly, less reliable predictions are expected for unknown inputs or inputs accumulating high rates of mutations. According to Bucher et al. [72], the accuracy of prediction and the predictivity score threshold levels are inversely proportional, or there is an increase in the number of the false positives as soon as the predictivity score threshold level starts to drop out [38]. Given that MEME primarily scans for un-gapped motifs, motifs containing indels (insertions/deletions) might be neglected. The number of input sequences, which is currently limited up to 50, is another limitation, in addition, to the MEME application. On the other hand, the TOMTOM program performs input motif query searching against a database of known motifs. Given the un-gapped alignment nature of the algorithm, motifs accumulating indels may not be recognized. In such cases, the method of Sandelin and Wasserman [73] would be more appropriate, as it preforms gapped motif-to-motif alignments.

5. Conclusions

This study highlights the importance of gaining a deeper understanding of the complex network of elements regulating the expression of male infertility-associated genes. By clarifying the presence of multiple promoters, identifying candidate transcription factor binding motifs, and revealing functional developments, we have shed light on the complex regulatory networks leading to male infertility. These findings not only deepen our understanding of the molecular mechanisms underlying male infertility, but also hold promise for advancing diagnostic approaches in this field. Future experimental validation of these computational predictions will be helpful in translating these insights into clinical applications, potentially helping the development of targeted therapies and personalized treatments for male infertility. We believe that our study will be a roadmap for further research in order to establish a rapid, individual, and detailed diagnosis of idiopathic infertility in couples as a cause of nonsyndromic male infertility.

Author Contributions: Conceptualization, D.H. and D.S.; methodology, D.H. and D.S.; software, D.H.; validation, D.S.; formal analysis, D.H.; investigation, D.H.; resources, D.H.; data curation, D.H.; writing—original draft preparation, D.H. and D.S.; writing—review and editing, D.H. and D.S.; visualization, D.H.; supervision, D.S.; project administration, D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, [D.H.], upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Venkatesh, T.; Suresh, P.S.; Tsutsumi, R. New insights into the genetic basis of infertility. *Appl. Clin. Genet.* **2014**, *7*, 235–243.
2. Agarwal, A.; Mulgund, A.; Hamada, A.; Chyatte, M.R. A unique view on male infertility around the globe. *Reprod. Biol. Endocrinol.* **2015**, *13*, 37. [CrossRef] [PubMed]
3. Shah, K.; Sivapalan, G.; Gibbons, N.; Tempest, H.; Griffin, D.K. The genetic basis of infertility. *Reproduction* **2003**, *126*, 13–25. [CrossRef]
4. Ferlin, A.; Raicu, F.; Gatta, V.; Zuccarello, D.; Palka, G.; Foresta, C. Male infertility: Role of genetic background. *Reprod. Biomed. Online* **2007**, *14*, 734–745. [CrossRef] [PubMed]
5. O'brien, K.L.F.; Varghese, A.C.; Agarwal, A. The genetic causes of male factor infertility: A review. *Fertil. Steril.* **2010**, *93*, 1–12.
6. Ferlin, A.; Arredi, B.; Foresta, C. Genetic causes of male infertility. *Reprod. Toxicol.* **2006**, *22*, 133–141. [CrossRef]
7. Jenkins, T.G.; Carrell, D.T. The sperm epigenome and potential implications for the developing embryo. *Reproduction* **2012**, *143*, 727–734. [CrossRef]
8. Chianese, C.; Gunning, A.C.; Giachini, C.; Daguin, F.; Balercia, G.; Ars, E.; Krausz, C. X chromosome-linked CNVs in male infertility: Discovery of overall duplication load and recurrent, patient-specific gains with potential clinical relevance. *PLoS ONE* **2014**, *9*, e97746. [CrossRef] [PubMed]
9. Krausz, C. Male infertility: Pathogenesis and clinical diagnosis. *Best Pract. Res. Clin. Endocrinol. Metab.* **2011**, *25*, 271–285. [CrossRef]
10. Tiepolo, L.; Zuffardi, O. Localization of factors controlling spermatogenesis in the nonfluorescent portion of the human Y chromosome long arm. *Hum. Genet.* **1976**, *34*, 119–124. [CrossRef]
11. Stojanov, D.; Koceski, S.; Mileva, A.; Koceska, N.; Bande, C.M. Towards computational improvement of DNA database indexing and short DNA query searching. *Biotechnol. Biotechnol. Equip.* **2014**, *28*, 958–967. [CrossRef] [PubMed]
12. Stojanov, D.; Madevska Bogdanova, A.; Orzechowski, T.M. TMO: Time and memory optimized algorithm applicable for more accurate alignment of trinucleotide repeat disorders associated genes. *Biotechnol. Biotechnol. Equip.* **2016**, *30*, 388–403. [CrossRef]
13. Stojanov, D.; Lazarova, E.; Veljkova, E.; Rubartelli, P.; Giacomini, M. Predicting the outcome of heart failure against chronic-ischemic heart disease in elderly population–Machine learning approach based on logistic regression, case to Villa Scassi hospital Genoa, Italy. *J. King Saud Univ.-Sci.* **2023**, *35*, 102573. [CrossRef]
14. Simoni, M.; Tempfer, C.B.; Destenaves, B.; Fauser, B.C.J.M. Functional genetic polymorphisms and female reproductive disorders: Part I: Polycystic ovary syndrome and ovarian response. *Hum. Reprod. Update* **2008**, *14*, 459–484. [CrossRef] [PubMed]
15. Yang, F.; Ouma, W.Z.; Li, W.; Doseff, A.I.; Grotewold, E. Establishing the Architecture of Plant Gene Regulatory Networks. In *Methods in Enzymology*; Academic Press: London, UK, 2016; Volume 576, pp. 251–304.
16. Lai, H.Y.; Zhang, Z.Y.; Su, Z.D.; Su, W.; Ding, H.; Chen, W.; Lin, H. iProEP: A computational predictor for predicting promoter. *Mol. Ther.-Nucleic Acids* **2019**, *17*, 337–346. [CrossRef] [PubMed]
17. Carvalho, A.M.; Freitas, A.T.; Oliveira, A.L.; Sagot, M.F. An efficient algorithm for the identification of structured motifs in DNA promoter sequences. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2006**, *3*, 126–140. [CrossRef] [PubMed]
18. Lambert, S.A.; Jolma, A.; Campitelli, L.F.; Das, P.K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T.R.; Weirauch, M.T. The human transcription factors. *Cell* **2018**, *172*, 650–665. [CrossRef] [PubMed]
19. Reményi, A.; Schöler, H.R.; Wilmanns, M. Combinatorial control of gene expression. *Nat. Struct. Mol. Biol.* **2004**, *11*, 812–815. [CrossRef]
20. Larson, D.; Bradford-Wilcox, J.; Young, L.S.; Sprague, K.U. A short 5′ flanking region containing conserved sequences is required for silkworm alanine tRNA gene activity. *Proc. Natl. Acad. Sci. USA* **1983**, *80*, 3416–3420. [CrossRef]
21. Morton, D.G.; Sprague, K.U. In vitro transcription of a silkworm 5S RNA gene requires an upstream signal. *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 5519–5522. [CrossRef]
22. Selker, E.U.; Morzycka-Wroblewska, E.; Stevens, J.N.; Metzenberg, R.L. An upstream signal is required for in vitro transcription of Neurospora 5S RNA genes. *Mol. Gen. Genet.* **1986**, *205*, 189–192. [CrossRef]
23. Garcia, A.D.; O'Connell, A.M.; Sharp, S.J. Formation of an active transcription complex in the Drosophila melanogaster 5S RNA gene is dependent on an upstream region. *Mol. Cell. Biol.* **1987**, *7*, 2046–2051. [PubMed]
24. Venkatesh, T.; Thankachan, S.; Kabekkodu, S.P.; Chakraborti, S.; Suresh, P.S. Emerging Patterns and Implications of Breast Cancer Epigenetics: An Update of the Current Knowledge. In *Epigenetics Reprod Health*; Academic Press: London, UK, 2021; Volume 21, pp. 295–324.
25. Mehmood, M.A.; Sehar, U.; Ahmad, N. Use of bioinformatics tools in different spheres of life sciences. *J. Data Min. Genom. Proteom.* **2014**, *5*, 1.
26. Stojanov, D. Structural implications of SARS-CoV-2 Surface Glycoprotein N501Y mutation within receptor-binding domain [499–505]–computational analysis of the most frequent Asn501 polar uncharged amino acid mutations. *Biotechnol. Biotechnol. Equip.* **2023**, *37*, 2206492. [CrossRef]
27. Stojanov, D. Phylogenicity of B. 1.1. 7 surface glycoprotein, novel distance function and first report of V90T missense mutation in SARS-CoV-2 surface glycoprotein. *Meta Gene* **2021**, *30*, 100967. [CrossRef] [PubMed]
28. Stojanov, D. Data on multiple SARS-CoV-2 surface glycoprotein alignments. *Data Brief* **2021**, *38*, 107414. [CrossRef] [PubMed]
29. Yahaya, T.O.; Liman, U.U.; Abdullahi, H.; Koko, Y.S.; Ribah, S.S.; Adamu, Z.; Abubakar, S. Genes predisposing to syndromic and nonsyndromic infertility: A narrative review. *Egypt. J. Med. Hum. Genet.* **2020**, *21*, 46. [CrossRef]

30. Guerri, G.; Maniscalchi, T.; Barati, S.; Gerli, S.; Di Renzo, G.C.; Della Morte, C.; Marceddu, G.; Casadei, A.; Laganà, A.S.; Sturla, D.; et al. Non-syndromic monogenic female infertility. *Acta Bio Medica Atenei Parm.* **2019**, *90*, 68.
31. Okutman, O.; Rhouma, M.B.; Benkhalifa, M.; Muller, J.; Viville, S. Genetic evaluation of patients with non-syndromic male infertility. *J. Assist. Reprod. Genet.* **2018**, *35*, 1939–1951. [CrossRef]
32. Joseph, S.; Mahale, S.D. Male Infertility Knowledgebase: Decoding the genetic and disease landscape. *Database* **2021**, *2021*, baab049.
33. Samuel, B.; Dinka, H. In silico analysis of the promoter region of olfactory receptors in cattle (Bos indicus) to understand its gene regulation. *Nucleosides Nucleotides Nucleic Acids* **2020**, *39*, 853–865. [CrossRef]
34. Beshir, J.A.; Kebede, M. In silico analysis of promoter regions and regulatory elements (motifs and CpG islands) of the genes encoding for alcohol production in *Saccharomyces cerevisiaea* S288C and Schizosaccharomyces pombe 972h. *J. Genet. Eng. Biotechnol.* **2021**, *19*, 8.
35. Bharathesree, R.; Murali, N.; Saravanan, R.; Anilkumar, R. Polymorphism of Keratin-Associated Protein (KAP) 6.1 gene and its association with wool traits of Sandyno and Nilagiri breeds of sheep. *Indian J. Anim. Res.* **2019**, *53*, 1566–1571. [CrossRef]
36. Bock, C.; Lengauer, T. Computational epigenetics. *Bioinformatics* **2008**, *24*, 1–10. [CrossRef]
37. Kanhere, A.; Bansal, M. Structural properties of promoters: Similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Res.* **2005**, *33*, 3165–3175. [CrossRef]
38. Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **1990**, *212*, 563–578. [CrossRef]
39. Michaloski, J.S.; Galante, P.A.; Nagai, M.H.; Armelin-Correa, L.; Chien, M.S.; Matsunami, H.; Malnic, B. Common promoter elements in odorant and vomeronasal receptor genes. *PLoS ONE* **2011**, *6*, e29065. [CrossRef]
40. Bailey, T.L.; Johnson, J.; Grant, C.E.; Noble, W.S. The MEME suite. *Nucleic Acids Res.* **2015**, *43*, W39–W49. [CrossRef]
41. Gupta, S.; Stamatoyannopoulos, J.A.; Bailey, T.L.; Noble, W.S. Quantifying similarity between motifs. *Genome Biol.* **2007**, *8*, R24. [CrossRef]
42. Buske, F.A.; Bodén, M.; Bauer, D.C.; Bailey, T.L. Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics* **2010**, *26*, 860–866. [CrossRef]
43. Takai, D.; Jones, P.A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 3740–3745. [CrossRef]
44. Zhao, Z.; Han, L. CpG islands: Algorithms and applications in methylation studies. *Biochem. Biophys. Res. Commun.* **2009**, *382*, 643–645. [CrossRef] [PubMed]
45. Hahn, M.A.; Wu, X.; Li, A.X.; Hahn, T.; Pfeifer, G.P. Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks. *PLoS ONE* **2011**, *6*, e18844. [CrossRef] [PubMed]
46. Wang, Q.; Xiong, F.; Wu, G.; Liu, W.; Chen, J.; Wang, B.; Chen, Y. Gene body methylation in cancer: Molecular mechanisms and clinical applications. *Clin. Epigenet.* **2022**, *14*, 154. [CrossRef] [PubMed]
47. Sandelin, A.; Carninci, P.; Lenhard, B.; Ponjavic, J.; Hayashizaki, Y.; Hume, D.A. Mammalian RNA polymerase II core promoters: Insights from genome-wide studies. *Nat. Rev. Genet.* **2007**, *8*, 424–436. [CrossRef] [PubMed]
48. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European molecular biology open software suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef] [PubMed]
49. Jedidi, I.; Ouchari, M.; Yin, Q. Autosomal single-gene disorders involved in human infertility. *Saudi J. Biol. Sci.* **2018**, *25*, 881–887. [CrossRef] [PubMed]
50. Zorrilla, M.; Yatsenko, A.N. The genetics of infertility: Current status of the field. *Curr. Genet. Med. Rep.* **2013**, *1*, 247–260. [CrossRef] [PubMed]
51. Smale, S.T.; Kadonaga, J.T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **2003**, *72*, 449–479. [CrossRef]
52. Juven-Gershon, T.; Hsu, J.Y.; Kadonaga, J.T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip. Rev. Dev. Biol.* **2006**, *1*, 40–51. [CrossRef]
53. Juven-Gershon, T.; Hsu, J.Y.; Theisen, J.W.; Kadonaga, J.T. The RNA polymerase II core promoter—The gateway to transcription. *Curr. Opin. Cell Biol.* **2008**, *20*, 253–259. [CrossRef]
54. Juven-Gershon, T.; Kadonaga, J.T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.* **2010**, *339*, 225–229. [CrossRef]
55. Djebali, S.; Davis, C.A.; Merkel, A.; Dobin, A.; Lassmann, T.; Mortazavi, A.; Tanzer, A.; Lagarde, J.; Lin, W.; Schlesinger, F.; et al. Landscape of transcription in human cells. *Nature* **2012**, *489*, 101–108. [CrossRef]
56. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **2014**, *507*, 462–470. [CrossRef]
57. de Klerk, E.; AC'tHoen, P. Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing. *Trends Genet.* **2015**, *31*, 128–139. [CrossRef]
58. Landry, J.R.; Mager, D.L.; Wilhelm, B.T. Complex controls: The role of alternative promoters in mammalian genomes. *Trends Genet.* **2003**, *19*, 640–648. [CrossRef]
59. FitzGerald, P.C.; Sturgill, D.; Shyakhtenko, A.; Oliver, B.; Vinson, C. Comparative genomics of Drosophila and human core promoters. *Genome Biol.* **2006**, *7*, R53. [CrossRef]

60. Hoskins, R.A.; Landolin, J.M.; Brown, J.B.; Sandler, J.E.; Takahashi, H.; Lassmann, T.; Yamamoto, J.-I.; Sekine, M.; Tsuritani, K.; Wakaguri, H.; et al. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res.* **2011**, *21*, 182–192. [CrossRef] [PubMed]
61. Rojas-Duran, M.F.; Gilbert, W.V. Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA* **2012**, *18*, 2299–2305. [CrossRef]
62. Kimura, K.; Wakamatsu, A.; Suzuki, Y.; Ota, T.; Nishikawa, T.; Yamashita, R.; Yamamoto, J.-I.; Sekine, M.; Tsuritani, K.; Wakaguri, H.; et al. Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **2006**, *16*, 55–65. [CrossRef]
63. Xu, C.; Park, J.K.; Zhang, J. Evidence that alternative transcriptional initiation is largely nonadaptive. *PLoS Biol.* **2019**, *17*, e3000197. [CrossRef] [PubMed]
64. Bantihun, G.; Kebede, M. In silico analysis of promoter region and regulatory elements of mitogenome co-expressed trn gene clusters encoding for bio-pesticide in entomopathogenic fungus, Metarhiziumanisopliae: Strain ME1. *J. Genet. Eng. Biotechnol.* **2021**, *19*, 94. [CrossRef] [PubMed]
65. Dai, Z.; Xiong, Y.; Dai, X. DNA signals at isoform promoters. *Sci. Rep.* **2016**, *6*, 28977. [CrossRef] [PubMed]
66. Xavier, M.J.; Salas-Huetos, A.; Oud, M.S.; Aston, K.I.; Veltman, J.A. Disease gene discovery in male infertility: Past, present and future. *Hum. Genet.* **2021**, *140*, 7–19. [CrossRef] [PubMed]
67. Lefebvre, V.; Dumitriu, B.; Penzo-Méndez, A.; Han, Y.; Pallavi, B. Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. *Int. J. Biochem. Cell Biol.* **2007**, *39*, 2195–2214. [CrossRef] [PubMed]
68. Hammes, A.; Guo, J.K.; Lutsch, G.; Leheste, J.R.; Landrock, D.; Ziegler, U.; Gubler, M.C.; Schedl, A. Two splice variants of the Wilms' tumor 1 gene have distinct functions during sex determination and nephron formation. *Cell* **2001**, *106*, 319–329. [CrossRef]
69. Heinlein, C.A.; Chang, C. Androgen receptor (AR) coregulators: An overview. *Endocr. Rev.* **2002**, *23*, 175–200. [CrossRef]
70. Parker, K.L.; Schimmer, B.P. Steroidogenic factor 1: A key determinant of endocrine development and function. *Endocr. Rev.* **1997**, *18*, 361–377. [CrossRef]
71. Uhlenhaut, N.H.; Jakob, S.; Anlag, K.; Eisenberger, T.; Sekido, R.; Kress, J.; Treier, A.-C.; Klugmann, C.; Klasen, C.; Holter, N.I.; et al. Somatic sex reprogramming of adult ovaries to testes by FOXL2 ablation. *Cell* **2009**, *139*, 1130–1142. [CrossRef]
72. Bucher, P.; Trifonov, E.N. Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res.* **1986**, *14*, 10009–10026. [CrossRef]
73. Sandelin, A.; Wasserman, W.W. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.* **2004**, *338*, 207–215. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.