



# IN SILICO INVESTIGATION OF GLOSSINA MORSITANS PROMOTERS



Sarah Wambui Mwangi

*Thesis submitted in fulfilment of the requirements for the degree of Doctor of  
Philosophy at the South African National Bioinformatics Institute, University of  
the Western Cape.*

Supervised by Prof. Alan G. Christoffels



# Abstract

Tsetse flies (*Glossina* spp) are the biological vectors for Trypanosomes, the causative agents of Human African Trypanosomiasis (HAT). HAT is a debilitating disease that continues to present a major public health problem and a key factor limiting rural development in vast regions of tropical Africa. To augment vector control efforts, the International *Glossina* Genome Initiative (IGGI) was established in 2004 with the ultimate goal of generating a fully annotated whole genome sequence for *Glossina morsitans*. A working draft genome of *Glossina morsitans* was available in 2011. In this thesis, transcriptional regulatory features in *Glossina morsitans* were analysed using the draft genome.

A method for TSS identification in the newly sequenced *Glossina morsitans* genome was developed using TSS-seq tags sampled from two developmental stages of *Glossina morsitans*. High throughput next generation sequencing reads obtained from *Glossina morsitans* larvae and pupae were used to locate transcription start sites (TSS) in the *Glossina morsitans* genome. TSS-seq tag clusters, defined as a minimum number of reads at the 5' predicted UTR or first coding exon, were used to define transcription start sites. A total of 3134 tag clusters were identified on the *Glossina* genome. Approximately 45.4% (1424) of the tag clusters mapped to the first coding exons or their proximal predicted 5'UTR regions and include 31 tag clusters that mapped to transposons. A total of 1101 (35.1%) tag clusters mapped outside the genic region and/or scaffolds without gene predictions and may correspond to previously un-annotated transcripts or noncoding RNA TSS.

The core promoter regions were classified as narrow or broad based on the number of TSS positions within a TSS-seq cluster. Majority (95%) of the core promoters analysed in this study were of the broad type while only 5% were of the narrow type. Comparison of canonical core promoter motif occurrences between random and *bona fide* core promoters showed that, generally, the number of motifs in biologically functional genomic windows in the true dataset exceeded those in the random dataset ( $p \leq 0.00164, 0.00135, 0.00185$  for the narrow, broad with peak and broad without peak categories respectively). Frequency of motif co-occurrence in core promoter was found to be fundamentally different across various initiation patterns. Narrow core promoters recorded higher frequency of the TATA-box and INR motifs and two-way motif co-occurrence showed that the TATA-box-INR pair is over-represented in the narrow category. Broad core promoters showed higher frequency of the BREd and MTE motifs and two-way motif co-occurrence showed that the MTE-DPE pair is over-represented in broad core promoters. TATA-less promoters account for 77% of the core promoters in this analysis. TATA-less core promoters showed a higher frequency of the MTE and INR motifs in contrast to observations in *Drosophila* where the DPE motif has been reported to occur frequently in TATA-less promoters. These motif combinations suggest their equal importance to transcription in their corresponding promoter classes in *Glossina morsitans*.

Nucleotide composition analysis showed that *Glossina morsitans* core promoters exhibit propensity for the AT dinucleotides unlike mammalian core promoters which displayed propensity for the CG dinucleotides. The variation in dinucleotide composition suggests a fundamental difference in global promoter architecture between mammals and insects.

A comparative genomics approach was employed to identify *Glossina morsitans* immunity genes using orthologous genes in *Drosophila melanogaster* and select blood-feeding insects. The evolutionary relationships between insect vector proteomes identified 190 putative immunity genes in *Glossina morsitans*. Essentially, majority of

*Glossina morsitans* immunity gene families were found to be systematically fewer than in other insect vectors. Most of the *Glossina* immunity genes with an experimentally verified TSS were found to be constituents of both developmental and immunity pathways. Proximal promoters of *Glossina morsitans* immunity genes were extracted and transcription factor binding site profiles established using an *ab initio* methodology. Transcription factor binding sites were compared with experimentally determined motifs from the JASPAR insect and vertebrate databases. Majority of the transcription factor binding sites on *Glossina morsitans* promoters of immunity genes were found to have experimental evidence implicating them as regulators of immunity and development such as apoptosis and autophagy. The Homeo-box class of transcription factors constituted majority of transcription factors identified as putative immune regulators of *Glossina morsitans* immunity.

Finally, proximal promoter regions of genes with experimentally verified TSS were extracted and compiled into a resource named GmPromDB using a MySQL relational database. PERL CGI was used for processing, and preparing HTTP requests and responses, while open source application programs HTML and CSS were also employed to design the front end. GBROWSE was embedded in the front end to facilitate viewing of the mapping profiles of TSS-seq reads on the genome. GmPromDB is a useful resource for *Glossina morsitans* promoter research and is accessible to the public use via a web interface (URL). The TSS locations are being integrated with the VectorBase resource at the European Bioinformatics Institute ([www.vectorbase.org](http://www.vectorbase.org)).

The work presented herein is a foundation to advance current understanding of the complex biological processes involved in *Glossina morsitans* transcriptional control mechanisms. The study has generated *in silico* inferred hypotheses that can be used to generate regulatory networks or be tested experimentally in subsequent studies.

# Keywords

*Glossina morsitans*

Human African Trypanosomiasis

Genome

TSS-seq

Transcription

Transcription start site

Promoter

Transcription regulation

Transcription factor binding site

Database



# Declaration

I, Sarah Wambui Mwangi, do hereby declare that "*In silico* investigation of *Glossina morsitans* promoters" is my own work. This work has not been presented in any university for examination. All resources I have used or quoted and all work which was the result of joint effort have been indicated and acknowledged by complete references.



Sarah Wambui Mwangi

Signed: .....

27th February 2014

# Acknowledgement

*Gracias a todos mucho!*

I want to acknowledge the producer Prof Alan Christoffels and production manager Dr Sumir Panji, for a brilliant casting of the crew. In a very special way, I would like to extend gratitude to my supporting actors, Mushal Allam, Ibrahim Ahmed, Christopher Mvelase, Darlington Mapiye, Zahra Jalali, Jean-Baka Domelevo and Siaka Lougue for invaluable assistance during some of the most challenging shoots. Many thanks to my fellow co-actor Monique for slaving away with my literature review. Your feedback was extremely helpful and highly appreciated.

I would also like to extend my heartfelt gratitude to Karyn Mergy, Xiaobei Zhao and Albin Sandelin for their ideas, though not on location, their advice was extremely helpful with the most challenging shoots.

I want to thank our technical coordinators Mario, Peter and Long and the location managers, Samantha, Maryam, Fungiwe and Ferial for managing all other stuff on location.

Thanks family! Dad and Mom, Paulina, Martha, Rachael, Zipporah, Bernard and Hannah. You truly were my pillars. And to little Micheal, Irene, Cynthia, Maria and Merlin-John, I love you all very much. I thank you for your constant belief that I could succeed. I dedicate my work to you and I hope you know how much I truly appreciate all you have done for me. Beyond doubt, I would not have been capable

of this if not for your constant encouragement.

Last and most importantly to my heavenly father. You gave beauty for ashes.

*Thank you all so very much!!!!*





*Remember Red, hope is a good thing, maybe the best of things, and no good thing  
ever dies.*

*-Andy Dufresne (in letter to Red), The Shawshank redemption.*



# Dedication...

*To my loving parents Joseph and Lucia Mwangi*

*At the end of the day, the most overwhelming key to a child's success is the positive involvement of parents. Jane D. Hull.*



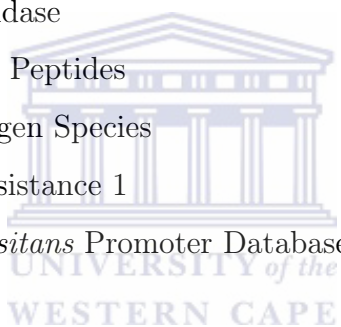
# Abbreviations

HAT	Human African Trypanosomiasis
VSG	Variable Surface Glycoprotein
ROS	Reactive Oxygen Species
DNA	Deoxyribonucleic acid
IGGI	International Glossina Genome Initiative
WGS	Whole genome shotgun
RNA	Ribonucleic acid
Mb	Megabases
EST	Expressed Sequence Tag
cDNA	Complementary DNA
mRNA	Messenger RNA
TSS	Transcription start site
RNAP	RNA Polymerase II
TFBS	Transcription Factor Binding Sites
GTFBS	General Transcription Factor Binding Sites
STFBS	Specific Transcription Factor Binding Sites
TF	Transcription Factor
TFIIB	Transcription Factor IIB
BREu	Upstream TFIIB Recognition Element
BREd	Downstream TFIIB Recognition Element
INR	Initiator
MTE	Motif Ten Element
DPE	Downstream Promoter Element



NGS	Next Generation Sequencing
CAGE	Capped Analysis of Gene Expression
PFM	Position Frequency Matrix
PWM	Position Weight Matrix
IUPAC	International Union of Pure and Applied Chemistry
MEF2	Myocyte Enhancer Factor-2
HNF4 $\alpha$	Hepatocyte Nuclear Factor 4 $\alpha$
Egr-1	Early growth response protein 1
CTF	CCAAT box-binding Transcription Factor
HMM	Hidden Markov Model
EM	Expectation maximization
MEME	Multiple Expectation Maximization for Motif Elicitation
OOPS	One Occurrence Per Sequence
SELEX	Systematic Evolution of Ligands by Exponential enrichment
ChIP	Chromatin Immunoprecipitation
DIP-chip	DNA immunoprecipitation
EMSA	Electrophoretic Mobility Shift Assays
ModENCODE	Model organism Encyclopedia of DNA Elements
TSR	Transcriptional Start Region
BAP	Bacterial Alkaline Phosphatase
TAP	Tobacco Acid Phosphatase
PCR	Polymerase Chain Reaction
TBP	TATA Binding Protein
TAF	TATA Associated Factors
PCR	Polymerase Chain Reaction
CDS	Coding Sequence

UTR	Untranslated Region
GO	Gene Ontology
PRR	Pattern Recognition Receptors
PAMPS	Pathogen Associated Molecular Patterns
LPS	Lipopolysaccharides
PGN	Peptidoglycan
PGRPS	Peptidoglycan Recognition Proteins
TEPS	Thioester-containing proteins
LIRMS	Leucine-rich immune proteins
CTLS	C-type lectins
PPO	Prophenol-oxidase
AMPS	Antimicrobial Peptides
ROS	Reactive Oxygen Species
OXR1	Oxidation Resistance 1
GmPromDB	<i>Glossina morsitans</i> Promoter Database



# Contents

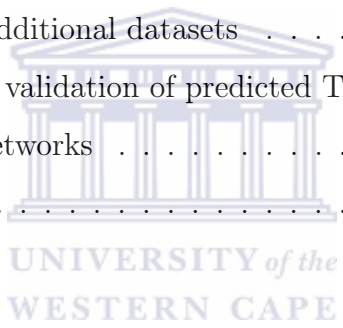
<b>1</b>	<b>Background and literature review.</b>	<b>1</b>
1.1	Tsetse flies and Trypanosomiasis . . . . .	1
1.2	Tsetse-Trypanosome interactions . . . . .	3
1.2.1	Trypanosome elimination strategies by Tsetse fly . . . . .	5
1.2.2	Survival tactics by Trypanosomes . . . . .	6
1.3	Control of Human African Trypanosomiasis . . . . .	6
1.4	The Tsetse fly genome: prospects of novel/improved vector control strategies . . . . .	7
1.4.1	<i>Glossina morsitans</i> genome sequencing effort and preliminary results . . . . .	10
1.4.2	Manual annotation of gene models . . . . .	12
1.5	Annotating transcriptional control elements . . . . .	13
1.5.1	RNA polII mediated basal transcription initiation . . . . .	13
1.5.2	A synopsis of basal transcriptional control modules (GTFBSs)	16
1.5.3	Emerging characteristics of core promoter architecture . . . . .	18
1.5.4	Evolution of core promoter architecture . . . . .	20
1.6	Strategies for TSS identification . . . . .	28
1.6.1	<i>In silico</i> prediction of promoters . . . . .	28
1.6.2	Experimental determination of promoters . . . . .	30
1.7	Strategies for identification of transcription factor binding sites . . . . .	31
1.7.1	Computational prediction of TFBSs . . . . .	31
1.7.2	Experimental determination of TFBSs . . . . .	42

1.8	The <i>Drosophila melanogaster</i> model for transcription regulation in insects . . . . .	44
1.9	Thesis rationale and objectives . . . . .	46
<b>2</b>	<b>Analysis of core promoter motifs in the genome of <i>Glossina morsitans</i> using TSS-seq.</b>	<b>49</b>
2.1	TSS-seq and promoter identification . . . . .	52
2.2	Properties of eukaryotic core promoters . . . . .	53
2.3	A summary of chapter objectives . . . . .	55
2.4	Methodology . . . . .	56
2.4.1	Acquisition and mapping of <i>G.morsitans</i> TSS-seq data . . . . .	56
2.4.2	TSS and promoter identification algorithm . . . . .	59
2.4.3	Analysis of core promoter properties . . . . .	64
2.4.4	Analysis of functional aspects of different promoter classes . . . . .	68
2.5	Results . . . . .	70
2.5.1	Genome mapping statistics . . . . .	70
2.5.2	Clustering statistics . . . . .	70
2.5.3	Delineation of promoter classes . . . . .	72
2.5.4	Core promoter extraction and classification . . . . .	75
2.5.5	Comparison of core promoter nucleotide distribution . . . . .	75
2.5.6	Annotation of core promoter motifs . . . . .	77
2.5.7	Functional classification of different promoter classes . . . . .	82
2.6	Discussion . . . . .	86
2.7	Conclusion, limitations and future work . . . . .	93
<b>3</b>	<b><i>In silico</i> analysis of promoters of <i>Glossina morsitans</i> immunity genes.</b>	<b>95</b>
3.1	Insect immunity: genes and pathways . . . . .	97
3.1.1	Pathogen recognition . . . . .	97
3.1.2	Signal transduction and modulation . . . . .	98
3.1.3	Pathogen elimination . . . . .	99
3.2	Characterization of promoters of insect immunity genes . . . . .	101

3.3	Promoters of immunity genes as a potential tool for design of novel vector control methods . . . . .	102
3.4	A summary of chapter objectives . . . . .	103
3.5	Methodology . . . . .	104
3.5.1	Compilation of insect proteomes and immunity genes . . . . .	104
3.5.2	Identification of immunity genes with a transcriptional signal . . . . .	105
3.5.3	Identification of immunity genes with possible developmental roles . . . . .	105
3.5.4	Promoter extraction . . . . .	105
3.5.5	Identification of transcription factor binding sites . . . . .	106
3.5.6	Analysis of TFBSs overrepresentation . . . . .	107
3.6	Results . . . . .	109
3.6.1	Comparison of the number of immunity genes between <i>G. morsitans</i> and other dipterans . . . . .	109
3.6.2	Promoter extraction and TFBSs analysis . . . . .	112
3.6.3	Predicted and experimentally verified transcription factor binding sites . . . . .	117
3.7	Discussion . . . . .	125
3.8	Conclusion, limitations and future work . . . . .	130
<b>4</b>	<b>GmPromDB: A database of <i>Glossina morsitans</i> promoters.</b>	<b>132</b>
4.1	Biological databases: a preamble . . . . .	134
4.2	Need for GmPromDB . . . . .	136
4.3	Data assembly . . . . .	136
4.4	Database design . . . . .	136
4.5	Database utility . . . . .	138
4.5.1	Home page . . . . .	138
4.5.2	Search page . . . . .	138
4.5.3	Search entry record . . . . .	139
4.5.4	Downloading promoters . . . . .	142



4.6	Conclusion, limitations and future work . . . . .	144
<b>5</b>	<b>Summary and Perspective</b>	<b>146</b>
5.1	Major contributions of this work . . . . .	147
5.1.1	Location of TSS in the <i>G.morsitans</i> genome . . . . .	147
5.1.2	Elucidation of <i>G.morsitans</i> core promoter architecture . . . . .	148
5.1.3	Promoter content of <i>G.morsitans</i> immunity genes . . . . .	148
5.1.4	A repository for <i>G.morsitans</i> promoters . . . . .	149
5.2	Perspective . . . . .	150
5.2.1	Improvement of the <i>G.morsitans</i> genome assembly and annotation . . . . .	150
5.2.2	Inclusion of additional datasets . . . . .	151
5.2.3	Experimental validation of predicted TFBSs . . . . .	152
5.2.4	Regulatory networks . . . . .	152
5.3	Final remarks . . . . .	153
	<b>Bibliography</b>	<b>154</b>
<b>6</b>	<b>Appendices</b>	<b>200</b>
6.1	<b>Appendix one: calculation of N50 genome statistics for the <i>G.morsitans</i> draft assembly . . . . .</b>	<b>201</b>
6.2	Appendix two: graphical representations of quality scores before and after trimming . . . . .	204
6.3	Appendix three: <i>G.morsitans</i> specific RNA POLII matrices and motif pictograms . . . . .	205
6.4	Appendix four: Comparison of the number of immunity genes between <i>G. morsitans</i> and other dipterans . . . . .	211



# List of Figures

1.1	The Trypanosomiasis transmission cycle presented as a triangle of interactions. . . . .	3
1.2	Life cycle of <i>Trypanosoma brucei</i> . . . . .	4
1.3	Organization of eukaryotic basal transcriptional control modules. . . . .	16
1.4	Narrow versus broad core promoters. . . . .	19
1.5	PFM for several Myocyte Enhancer Factor-2 (MEF2) TFBSs. . . . .	32
1.6	Sequence logo depicting the information content of MEF2. . . . .	36
1.7	Representation of the markov chain concept. . . . .	37
1.8	Representation of the hidden markov concept. . . . .	38
1.9	The interplay between computational and experimental techniques. . . . .	43
1.10	Illustration of thesis contents . . . . .	48
2.1	Outline of the TSS-seq procedure. . . . .	52
2.2	Trends of the quality filtering and mapping procedures. . . . .	57
2.3	Representation of the classification of 5'UTR tag clusters. . . . .	60
2.4	Workflows summarizing the mapping and clustering protocols. . . . .	61
2.5	Histogram of a sample TSS-seq tag cluster. . . . .	63
2.6	Summary of core promoter motif analysis methodology. . . . .	68
2.7	Summary of core promoter GO annotations analysis methodology. . . . .	69
2.8	Relationship between <i>Sg</i> value and number of TSS positions . . . . .	72
2.9	Relationship between number of TSS positions. . . . .	73
2.10	Graphical impressions of representative tag clusters for the various promoter classes. . . . .	74

2.11	Combined promoter nucleotide composition graphs. . . . .	76
2.12	<i>G.morsitans</i> core promoter nucleotide frequency surrounding the TSS. . . . .	77
2.13	Graphical summary of core promoter instances in the various promoter classes. . . . .	79
2.14	Graphical summary of core promoter instances in the TATA containing and TATA-less core promoters. . . . .	80
2.15	Ontology terms occurring in the 75th percentile of narrow core promoters. . . . .	83
2.16	Ontology terms occurring in the 75th percentile of broad with peak core promoters . . . . .	84
2.17	Ontology terms occurring in the 75th percentile of broad without peak core promoters. . . . .	85
3.1	Generalized insect innate immune pathways based on <i>Drosophila</i> literature. . . . .	99
3.2	Protocols and tools used for analysis of promoters of <i>G.morsitans</i> immunity genes. . . . .	108
3.3	Distribution of organism PWM instances that are over-identified in <i>G.morsitans</i> promoters of immunity genes. . . . .	115
3.4	Summary of transcription factor families characterised in <i>G.morsitans</i> proximal promoters. . . . .	116
3.5	Graphical representation of TF IDs with high frequency. . . . .	117
4.1	The conventional three-tier database architecture. . . . .	135
4.2	Simplified representation of the database design. . . . .	137
4.3	A snapshot of GmPromDB's home page. . . . .	138
4.4	A snapshot of the search page for GmPromDB. . . . .	139
4.5	A snapshot of the search entry record. . . . .	140
4.6	A snapshot of the search entry record. . . . .	141
4.7	A snapshot of downloads section. . . . .	142
4.8	A scaffold entry depicting all the genes in the scaffold. . . . .	143

- 6.1 Quality chart of the TSS-seq reads before (i) and after (ii) quality control.204
- 6.2 A graphical summary of the immunity gene families in selected dipterans.211



# List of Tables

1.1	The <i>G.morsitans</i> genome contig and scaffold assembly statistics . . .	10
1.2	Gene parameters in the <i>G.morsitans</i> genome . . . . .	11
1.3	Scaffold gene distribution in the <i>G.morsitans</i> genome . . . . .	12
2.1	Summary of read mapping statistics . . . . .	58
2.2	Core promoter genomic windows used for the analysis . . . . .	65
2.3	Summary of genome mapping statistics . . . . .	70
2.4	Summary of tag clustering statistics . . . . .	71
2.5	Summary of tag cluster types . . . . .	75
2.6	Comparison of core promoter motif instances between true and random datasets . . . . .	78
2.7	Two way motif co-occurrences . . . . .	81
2.8	Three way motif co-occurrences . . . . .	82
3.1	A summary the total number of protein coding genes for selected insect vectors . . . . .	104
3.2	A summary of the whole complement of immunity genes in selected insect vectors . . . . .	110
3.3	Summary of immunity genes with experimentally verified TSS . . . .	113
3.4	Predicted transcription factor binding site for <i>G.morsitans</i> immunity genes . . . . .	118

# Chapter 1

## Background and literature review.

### 1.1 Tsetse flies and Trypanosomiasis

The obligate blood feeders of the Hippoboscoidea superfamily are made up of three families, the *Glossinidae* (Tsetse flies), the *Hippoboscidae* (Louse flies) and *Nycteribiidae* (bat flies) [1]. Within the family *Glossinidae*, the genus *Glossina* (Tsetse flies) is placed as the sole member by most classifications. This genus includes up to thirty-four species and sub-species and is usually split into three major groups based on a combination of geographical, behavioral, genetic and physical traits. The three groups include; Savannah flies – sub-genus *Morsitans*, Riverine flies – sub-genus *Palpalis* and Forest flies – sub-genus *Fusca*. Nine of the thirty-four known species and sub-species of *Glossina*, belonging to either the *Palpalis* or *Morsitans* sub-genera are biological vectors for Trypanosomes.

During a blood meal, Tsetse flies lacerate the skin ingesting the blood and injecting saliva into the bite site. Tsetse saliva constitutes anti-clotting compounds and may also contain infective Trypanosomes. Infective Trypanosomes are the causative agents of Human African Trypanosomiasis (HAT)/ sleeping sickness, a debilitating disease that is lethal if it remains untreated. The particular ecological environment of Tsetse flies and Trypanosomes is such that the disease is only found in the inter-tropical regions of Africa [2]. *Glossina* spp can infect a wide range of hosts including

domesticated animals, thus HAT is persistent in agrarian populations. Whilst the sub-genus *Morsitans* flies portray zoophilic tendencies, the sub-genus *Palpalis* is generally anthropic and is adapted to domestic environments. *Glossina Palpalis* species are therefore vectors for the human-infective Trypanosomes whereas *Glossina morsitans* species are vectors for animal-infective Trypanosomes.

HAT continues to present a major public health problem and a key factor limiting rural development in vast regions of tropical Africa. Several epidemics struck parts of Africa in the beginning of the 20th century. Constant surveillance markedly reduced the disease such that by the mid 1960's the disease had almost disappeared. There had been a resurgence of trypanosomiasis in the last few decades due to lack of control programs within the endemic countries but concerted efforts pioneered by the world health organisation have seen a decline in the reportage of new cases to below 10,000 cases annually in the recent past [3]. Nonetheless, it is believed that this may be a gross underestimation due to the undeveloped health system infrastructure in the areas afflicted by HAT. Among protozoan infections in sub-Saharan Africa, HAT is ranked as one of the major parasitic disease affecting human health [4].

From a population perspective, persistence of HAT depends on a complex interplay between the three interacting organisms: the mammalian host, the parasitic trypanosome and the insect vector (Figure 1.1). Epidemiological modeling of these three interactions and their consequences for disease prevalence has established a strong link to socio-economic and political factors as well as understanding of the host-parasite interactions [5].

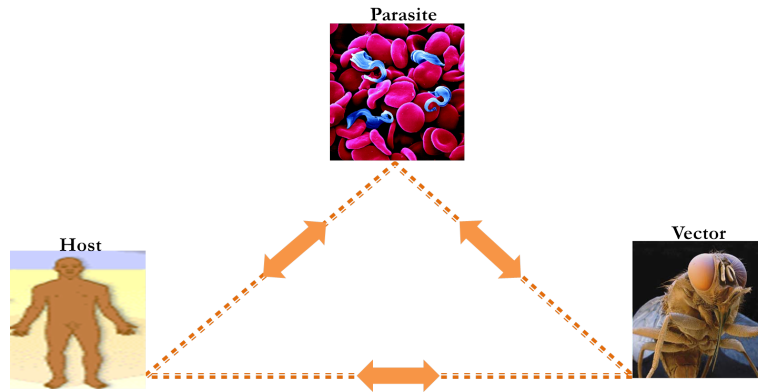


Figure 1.1: The Trypanosomiasis transmission cycle presented as a triangle of interactions. The nature of interactions is such that both the vector and the hosts are vital for production of a fully-fledged infection. Interruption of either the host-parasite or vector-parasite interactions would reduce the rate of disease transmission significantly.

## 1.2 Tsetse-Trypanosome interactions

The Trypanosome life cycle begins in the fly and is characterized by a succession of two stages: (i) the growing stage that occurs within the fly and is adapted to infection and (ii) the non-growing stage that occurs within the host bloodstream and is adapted to transmission. As the Trypanosomes change their environment, they exhibit gross change in morphology, an event that is coupled with adaptive activation and repression of metabolic pathways.

During a blood meal, the meal encased in the peritrophic membrane by the proventricular valve enters the midgut where establishment of ingested parasites takes place. While in the mammalian bloodstream, Trypanosomes exhibit extracellular “non-replicative” characteristics and are referred to as trypomastigotes. In the midgut, trypomastigotes multiply exponentially whilst transforming to procyclic (short-stumpy) forms. Procyclic trypomastigotes migrate from the midgut to the proventriculus where they proceed with replication and transform into epimastigotes. Epimastig-



otes proceed to the salivary glands where they stop replication to a non-dividing and infective form known as the metacyclic (long-slender) trypomastigotes [6] (see figure 1.2). In the salivary glands, metacyclic trypomastigotes initiate expression of a variable surface glycoprotein (VSG) coat as a pre-adaptation immune evasion in the host bloodstream [7]. The period from ingesting infected blood to appearance of infective metacyclic trypomastigotes varies from one to three weeks. However, a fly remains infective for life. Additionally, the whole infective cycle is probably completed successfully in only one for every ten infected flies [8].

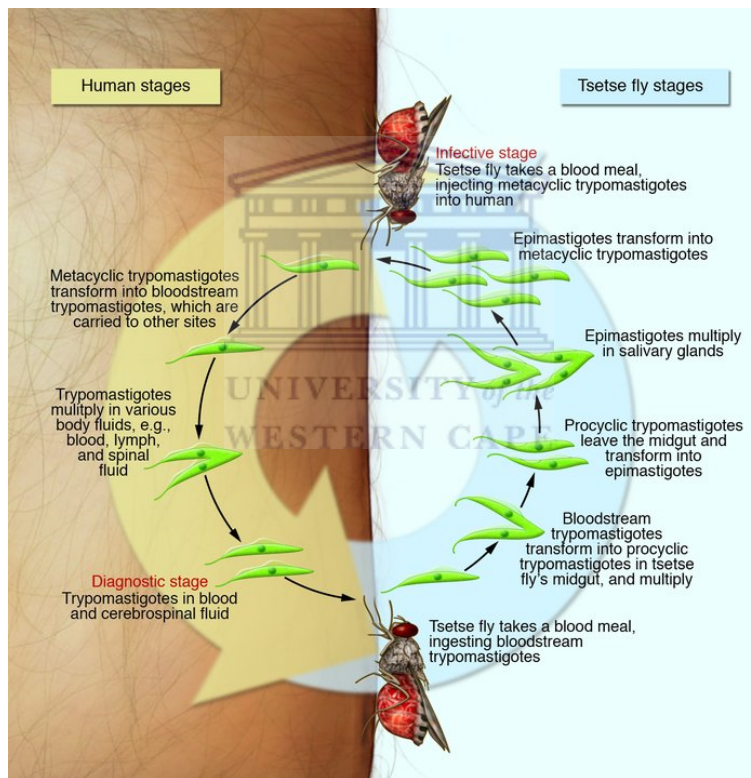


Figure 1.2: Life cycle of *Trypanosoma brucei*. The life cycle begins in the fly and matures in the mammalian host [9].

The elaborate events during the parasite life cycle dictate that, for successful transmission, accurate crosstalk between the Trypanosome and the Tsetse fly is paramount. It is this crosstalk that facilitates uninterrupted differentiation, multiplication and finally migration by the parasite. Ultimately, these events lead to infection of a new host [10]. Tsetse flies and Trypanosomes have co-evolved alongside each other

to ensure mutual survival. On the one hand, Tsetse flies have evolved mechanisms to protect them from succumbing to invasion by Trypanosomes (see section 1.2.1), whereas Trypanosomes have evolved strategies for surviving the anti-Trypanosome episodes in the fly to ensure maturation to mammalian infective forms in the salivary glands [10] (see section 1.2.2).

### 1.2.1 Trypanosome elimination strategies by Tsetse fly

Trypanosome elimination strategies by Tsetse fly occurs via a series of events that can be divided into two phases: (i) establishment of Trypanosomes as a dividing procyclic population in the Tsetse midgut and (ii) maturation into human infective forms in the salivary glands. During establishment in the midgut, most Trypanosomes are eliminated by both environmental and Tsetse-intrinsic factors which include: (i) midgut lectins that are capable of killing Trypanosomes by an apoptosis-like process [11], (ii) competition with other trypanosome strains, in the case of a mixed infection and (iii) a robust immune response that Tsetse flies mount against invading pathogens. The immune response is believed to be the chief mechanism through which Tsetse flies eliminate Trypanosomes from the midgut [12–15].

Not all Trypanosomes that evade the fly immune arsenal mature to infective forms. Male flies show a higher rate of Trypanosome maturation. The underlying mechanism of these sex differences was hypothesized to be the operation of a product(s) of an X-linked gene that kills or prevents migrating parasites from maturing [16]. In addition, Trypanosome strain was implicated in maturation, where the probability of maturation was significantly greater for Trypanosomes that are human serum sensitive than for those that are human serum-resistant [17].

### 1.2.2 Survival tactics by Trypanosomes

Parts of the by-products generated during catabolism of a blood meal are toxic reactive oxygen species (ROS) that are coincidentally released as part of the immune arsenal. To counter the effect of the blood meal ROS, Tsetse flies release antioxidants in a process that turns out to be beneficial for the developing Trypanosomes as the antioxidants reduce the midgut environment protecting Trypanosomes from cell death induced by ROS. By utilizing direct *in vivo* feeding experiments, MacLeod and colleagues [18] showed that maturation of *Trypanosoma brucei brucei* infections in Tsetse midgut is promoted by a range of antioxidants.

During maturation to metacyclic forms in the salivary gland, the Trypanosome's flagellum develops extraordinary branched outgrowths which are attached to the salivary gland microvilli [19]. The nature of the attachment of the flagellum to the microvillus membrane is difficult to study *in vitro*. Nonetheless, inhibition of attachment does not prevent division of the Trypanosomes, but it does prevent their differentiation into metacyclic forms. Attachment therefore appears to have a developmental significance and to constitute an essential part of the program for trypanosome maturation to infective forms [19]. Trypanosomes may promote their transmission through manipulation of the Tsetse feeding behavior by modifying the saliva composition. Tsetse flies with salivary gland infections display a significantly prolonged feeding time relative to the non-infected counterparts. This prolonged feeding enhances the likelihood of transmitting infective Trypanosomes to many hosts during a single blood meal cycle [20].

## 1.3 Control of Human African Trypanosomiasis

To realize significant reduction in HAT transmission, sustainable control would utilise an integrated approach that exploits both host-Trypanosome and Tsetse-Trypanosome interactions. The host-Trypanosome interactions have been primarily exploited by use

of chemotherapy treatment. Only four drugs are available for the chemotherapy of HAT; Eflornithine, Melarsoprol, Pentamidine, and Suramin [21]. In addition to the paucity of clinically approved drugs, treatment is complicated by the fact that different drug combinations have to be used for various Trypanosome subspecies as well as distinct developmental stages [22]. Efforts to develop a vaccine for immunization have been hampered by the parasites ability to express an ever changing variable surface glycoprotein (VSG) during host invasion. Rapid rate of switching the VSGs generates diversity in parasite population facilitating immune evasion. Evasion of the immune arsenal results in continuous persistence of parasites in the host, increasing the chances for transmission [23]. Acquired immunity is seldom completely protective against Trypanosome reinfection and its effectiveness seems to be determined by the duration and intensity of previous exposure to infection [24]. Until recently, one of the leading difficulties in anti-Trypanosome drug development was the inability to identify suitable targets from a small percentage of the whole complement of Trypanosome genes that were known [25]. Massive amounts of information accrued from Trypanosome genome sequencing and analyses have facilitated several new approaches for drug target identification [25] [26]. Furthermore, post genome follow-up studies are continuously garnering comprehensive and accurate target identification using an array of *in-silico* approaches. For example, Amaro and colleagues [27] identified drug-like inhibitors of an crucial RNA-editing ligase in *Trypanosoma brucei* by employing *in silico* approaches.

## 1.4 The Tsetse fly genome: prospects of novel/improved vector control strategies

Information accrued from genomic analyses may well facilitate design of new schemes for blocking HAT transmission through improved and/or novel vector control strategies. For instance, comparative analyses of the genetic basis of Tsetse-Trypanosome

interactions will undoubtedly yield important advancements in the study of Tsetse-Trypanosome co-evolution. Such studies may reveal novel opportunities for vector control such as methods targeting the inhibition of blood feeding or interfering with the Trypanosome development. In addition, data from genome analyses could be used to develop novel DNA-based diagnostic tools for characterizing vectors and identification of potential targets for insecticides [28].

Vector transgenesis and paratransgenesis utilize genomic information to reduce insect vectorial capacity. These techniques rely on direct genetic manipulation of disease vectors with an aim of reducing their ability to facilitate parasite development. Despite the promise that vector transgenesis holds, the viviparous nature of the Tsetse fly unlike other insect vectors is limiting because the reproductive ability of the female Tsetse fly is approximately five to eight offsprings for an entire lifetime [29]. During paratransgenesis, genetically modified vector symbionts are utilized to express molecules within the vector that interfere with parasite development. In Tsetse flies, transformation of the maternally inherited symbionts (*Sodalis glossinidae*) to express trypanocidal products constitutively using genetic methods has been proposed as a novel technique of vector control [30]. When microinjected into the haemolymph of the female parent, recombinant *Sodalis* species were successfully acquired by progeny flies and were propagated in subsequent generations with high fidelity [31]. Members of *Sodalis glossinidae* are localized within the midgut where Trypanosome establishment takes place, thus expression of trypanocidal products by *Sodalis glossinidae* can potentially block parasite establishment. In addition, *Sodalis glossinidae* has co evolved with the fly's immune system to display a high level of resistance to the insect's immune arsenal [32]. A recent study by De Vooght and colleagues [33] has reinforced the notion applicability of paratransgenesis by showing that functional Trypanosome-interfering nanobodies in *Sodalis glossinidus* can be expressed and released extracellularly.

The genomes of two players in the HAT transmission cycle have been completed: see

Venter *et al.*, [34] for the human genome, Berriman *et al.*, [25] for the *Trypanosoma brucei brucei* genome and Jackson *et al.*, [26] for the *Trypanosoma brucei gambiense* genome. The completed genomes of the African Trypanosomes have provided information that could be valuable in control of HAT. For example, the *Trypanosoma brucei brucei* genome provided insights into novel drug targets including; (i) The glycosylphosphatidylinositol anchors that facilitate attachment of components of the variable surface glycoprotein coat and (ii) The isoprenoid metabolism pathway which is susceptible to antifungal agents [25]. On the other hand, the *Trypanosoma brucei gambiense* genome provided the first estimate of intraspecific genomic variation within *Trypanosoma brucei*. Analysis of the VSG domains showed that the fundamental structural components of VSG domains are conserved across the two sub-species [26].

The International *Glossina* Genome Initiative (IGGI) was established in 2004 with the ultimate goal of generating a fully annotated whole genome sequence for *G. morsitans*. This is expected to improve understanding of vectorial capacity, as well as the utilisation of that understanding to halt the HAT transmission cycle. Currently, a working draft genome of *G. morsitans* is available at VectorBase [35]. Forging a link among the genomic sequences through comparative analysis of the closely related species may for example, generate testable hypotheses of genes that might be responsible for differences in vector competence. Hereof, a proposal to sequence five additional *Glossina* genomes including the vectors for human infective Trypanosomes, *G. palpalis* has been put forward. Additional *Glossina* species to be sequenced include *G. fuscipes*, *G. pallidipes*, *G. brevipalpis* and *G. austeni* [28]. Variations in genomic regions among *Glossina* spp can be rapidly identified and evaluated for relevance in the development of novel strategies to stop the transmission cycle.

### 1.4.1 *Glossina morsitans* genome sequencing effort and preliminary results

Whole genome shotgun (WGS) sequencing entails shearing of organismal DNA into fragments of several thousand nucleotides. These fragments are cloned directly into a plasmid vector after which sequencing is done such that each base pair is covered several times. Sequence data from each end of the cloned inserts are known as mate pairs and they are assembled to reconstruct the complete genome. Assembly results in a set of contiguous sequences referred to as contigs. Contigs are ordered and oriented such that the gaps between adjacent contigs are of known size and are spanned by clones with end sequences flanking the gap. The process of ordering contigs is what is referred to as scaffolding and it generates longer sequences commonly referred to as scaffolds. Gaps within scaffolds are known as ‘sequence gaps’ while gaps between scaffolds are known as ‘physical gaps’ as there are no clones identified spanning the gaps [36].

The *G.morsitans* genome has been sequenced and assembled at the Sanger institute using a combination of WGS and new Illumina RNA-seq technologies. The assembly statistics are presented in the table below;

Table 1.1: The *G.morsitans* genome contig and scaffold assembly statistics

Parameter	Contigs	Scaffolds
Total number	24,072	13,807
Total length*(1)	363 Mb	366 Mb
Average size	15,084 bp	26,522 bp
Longest	538,224 bp	25,362,821bp
N50 statistic*(2)	49,769 bp	120,413bp

\*(1) The *G.morsitans* genome is estimated as approximately 389 Mb [28] and the total number of gaps is roughly 1% excluding masking. Thus, approximately 99% of

the genome has been sequenced.

\*(2) N50 contig or scaffold size defines the size above which 50% of the assembly is found. For an explanation of the calculation of the N50 statistic, see appendix one.

Parallel to the *G.morsitans* genome sequencing effort, the IGGI consortium has generated expressed sequence reads (ESTs) and more recently transcript fragments using RNA-seq [37] from various tissues and developmental stages of *G.morsitans*. This was used to assist with the assembly and annotation of the genome. The genome was repeat-masked prior to annotation with known *Glossina* and *Drosophila* repeats from GenBank and with *de novo* repeats identified by running the REPEATMOD-ELER [38] program. A preliminary gene set generated using the ensembl gene-build pipeline [39] constitutes 12,220 protein-coding genes where alternatively spliced transcripts for 142 of these protein-coding genes were estimated. There are 64,464 exons predicted for the genome, an average of 5 exons per gene. Thus far, no data for coding and non-coding RNA genes has been available. The parameters for protein-coding genes are summarized below;

Table 1.2: Gene parameters in the *G.morsitans* genome

Description	Count
Number of protein-coding genes	24,072
Protein-coding genes with alternative transcripts	142*
Length of genome in protein-coding exons	20,798,601
Total number of exons	64,464
Length of genome in introns	75,928,103
Number of genes with 5' UTR	6937
Average (median) 5' UTR length	301
Number of genes with 3' UTR	7463
Average (median) 3' UTR length	651



\*130 protein-coding genes have two isoforms, eleven protein-coding genes have three isoforms and one protein-coding gene has four isoforms.

3057 out of 13807 of the assembled scaffolds contain protein-coding genes whose distribution is presented by the table below;

Table 1.3: Scaffold gene distribution in the *G.morsitans* genome

Ranges of gene count	Number of genes
Equal to 1	1255
2 to 5	1248
6 to 10	381
11 to 50	162
51 to 100	8
Greater than 100	4

Majority of the scaffolds ( 80%) contain less than five genes. Only twelve scaffolds may be considered ‘gene-rich’ as they harbor at least fifty genes. The genome is riddled with repeat elements. Preliminary calculations show that repeat regions constitute approximately one third of the genome predominantly in the gene poor regions.

## 1.4.2 Manual annotation of gene models

The initial phase of the genome annotation essentially identifies genes to reveal the functional significance of specific sections of the genomic sequence. Ordinarily, this phase incorporates three complementary approaches: (i) Computer programs are designed to perform *ab initio* gene prediction using cues provided by protein homologies, (ii) mRNAs are aligned to the assembled genomic sequence to reveal known genes and (iii) additional genes are found based on alignment of cDNAs to the assembled genomic sequence. Biological expertise is needed to improve the fidelity of the gene

models predicted using computational programs. Regarding the *G.morsitans* genome, gene models predicted by the Ensembl gene-build pipeline were presented to the annotators by means of a web portal [40] to enable the download of individual gene models together with their corresponding data such as cDNA and peptide sequences and genomic locations. Manual curation essentially involved structural modification of gene models (provided it was necessary) and assignment of gene names and symbols together with any other biological information that would improve the fidelity of the computational prediction. The manual curation process is still on-going.

## 1.5 Annotating transcriptional control elements

Even though gene prediction constitutes a bulk of the initial annotation process, some notion of function beyond what can be determined by gene prediction is understanding when and where a gene is expressed. Mechanisms controlling gene expression constitute regulatory events occurring at transcriptional, translational and post-translational levels. For a given cell, regulation of gene expression at the transcriptional level is perhaps the most important basis of cellular function. In essence, gene transcription exerts fundamental control over the abundance of nearly all of a cell's functional macromolecules by modulating and synchronizing multiple genes encoding products with interdependent activities. Regulation at the transcription level is governed by two main interacting systems: (i) marking of histone proteins on which DNA is wound with chemical tags to determine active or silent genomic regions (DNA methylation and chromatin remodeling) [41] and (ii) transcription factors that bind to DNA in promoters of genes to initiate transcription (Figure 1.3).

### 1.5.1 RNA polII mediated basal transcription initiation

The RNA polymerase II core promoter is simply defined as the sequence that guides transcription initiation. However, this modest definition belies a diverse, complex and

intricate basal transcriptional initiation program [42]. A promoter is the segment of DNA usually occurring upstream from a gene-encoding region. Positions in the promoter are designated relative to the transcriptional start site (TSS). The TSS is the first nucleotide of an RNA transcript where RNA polymerase binds to initiate transcription. Promoters lack universal structural features implying that no consistent sequence motifs exist for protein-coding genes. However, two functional features are constantly present though they cannot always be discerned from sequence information [43]. This includes the core promoter and the proximal promoter.

The core promoter consists of the minimal portion of the promoter required to properly initiate transcription. Typically, the core promoter encompasses the TSS and extends either upstream or downstream for additional 40-50 nucleotides [44]. The core promoter harbors the TSS and transcription factor binding sites (TFBSs). TFBSs are short degenerate motifs of approximately 5-15 nucleotides. It is at the core promoter that the basal transcription machinery is recruited and is composed of several transcription factors (TFs) that bind the TFBSs. A TF can be defined as a protein that binds to specific DNA sequences (motifs), thereby modulating gene transcription by acting as activators or repressors [45]. Core promoter TFBSs are also referred to as general TFBSs (GTFBS) as they are conserved among eukaryotic core promoters.

The proximal promoter is situated adjacent to the core promoter and extends approximately 250 nucleotides upstream of the TSS. It contains a collection of TFBSs that are also referred to as proximal/specific transcription factor binding sites (STFBSs) (Figure 1.3). Proximal TFBSs are a collection of diverse TFBSs that confer transcription specificity. Since the core promoter cannot generate functionally significant levels of mRNA singly [46] [47], proximal TFBSs and their corresponding TFs augment basal transcription initiation by generating an amplified response, to initiate gene transcription in a spatio-temporal mode.

Enhancers, silencers and insulators are additional components of the RNAP mediated basal transcription initiation. Enhancers are short regions of genomic DNA that are usually composed TFBSs groups working together enhance gene transcription [48]. Their position varies among genes; while some are close to the genes they act on others may be located on a different chromosome or within the intron of the gene that they regulate [49] [50]. Instead of acting on the promoter region itself, enhancers are bound by activator proteins that interact with a mediator complex. It is this mediator complex is responsible for recruiting RNAP and the general TFs to initiate transcription [49].

Silencers repress transcription [51]. Silencers share some of the properties of enhancers such as function independence of orientation and distance from the promoter [52]. Silencers may be situated as part of a distal enhancer or a proximal promoter or as an independent distal regulatory module in the target gene's intron or 3' UTR [48].

Insulators act as a barrier from the transcriptional activity of neighboring genes by limiting the action of transcriptional regulatory elements to defined regions. It is assumed that insulators exert function by blocking enhancer-promoter communication [48].

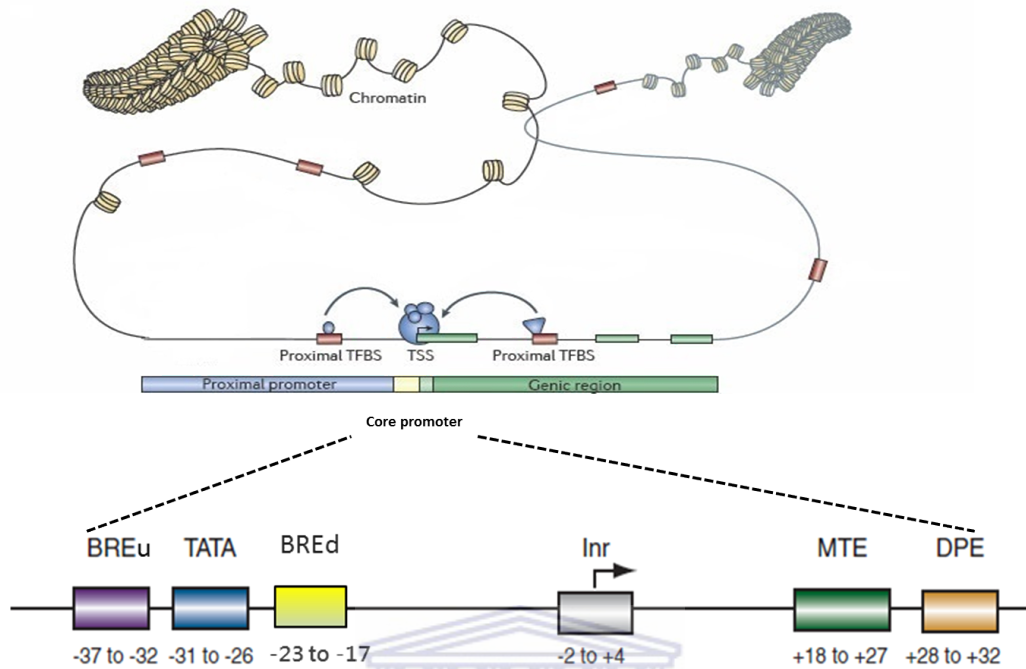


Figure 1.3: Organization of eukaryotic basal transcriptional control modules. Nucleic acid material wrapped around histones constitutes chromatin that may be silent (tightly wrapped) or active (accessible). The TSS is flanked by the core promoter. Core promoter motifs include the TFIIB recognition elements (BREu and BREd), TATA-box, initiator (INR), motif ten element (MTE) and the downstream promoter element (DPE). The BREu and BREd are an extension of a subset of TATA-boxes [48] [53].

### 1.5.2 A synopsis of basal transcriptional control modules (GTF-BSs)

Core promoter motifs serve as the docking site of the basal transcription machinery thus playing a crucial role in determining the position and directing the rate of transcription initiation [54]. Though variations exist, most core promoter motifs are conserved among eukaryotic core promoters. Each of these motifs may be found in a subset of core promoters. In principal, a core promoter will contain a combination of but rarely all motifs.

### 1.5.2.1 TATA-box and BRE

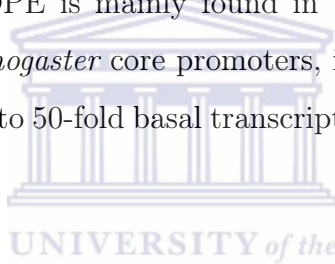
The TATA-box has the core DNA sequence TATAWAAR (degenerate nucleotides are designated according to the IUPAC code). It is often located within 25-35 nucleotides upstream of the TSS. This sequence binds a TATA binding protein (TBP), a TF that enables the formation of the RNA polymerase transcriptional complex by facilitating easy unwinding of DNA. The major core promoter-binding factor is TFIID. TFIID is a large complex comprising of the TBP and 13 associated factors known as TFIID associated factors (TAF). The TATA-box is directly recognized and bound by TBP while the TAFs interact with sequences upstream or downstream of the TATA-box [55] [56] [59–64]. The TATA box occurs in approximately 10-30% of all genes within a genome [44] [59] [65–67]. The BRE is a TFIIB binding site that is located immediately upstream (BREu) or downstream (BREd) of some TATA-boxes [68]. The BREu consensus is SSRCGCC [69] while the BREd consensus is RTDKKKK [63] (degenerate nucleotides are designated according to the IUPAC code). The BRE motifs activate or repress transcription depending on the promoter content

### 1.5.2.2 Initiator (INR)

The INR box is a pyrimidine-rich sequence that encompasses the TSS and the region immediately surrounding it. The consensus for the INR is YYANWYY in humans and TCAKY in *D. melanogaster* [70] (degenerate nucleotides are designated according to the IUPAC code). The INR motif is found in both TATA-containing as well as TATA-less core promoters [71]. Recent studies suggest that the INR consensus is composed of only YR, where R corresponds to the TSS [72] [73]. Even though some promoters lack an INR and TATA-box, only a limited number of characterized core elements function independently of INR and TATA [63].

### 1.5.2.3 Downstream core promoter element (DPE) and motif ten element (MTE)

The MTE was described as a new core promoter element for transcription by RNA polymeraseII by Lim and colleagues [74] in *D. melanogaster*. The MTE is located at positions +18 to +29 in the core promoter and is represented by the consensus sequence CSARCSSAACGS (degenerate nucleotides are designated according to the IUPAC code). The MTE requires precise INR-MTE spacing for transcription and can compensate for the loss of a DPE as well as the loss of a TATA-box [74]. The DPE is located precisely at +28 to +32 relative to the TSS position in the INR, and its consensus sequence is RGWYV [74] (degenerate nucleotides are designated according to the IUPAC code). The DPE is mainly found in TATA-less promoters. In the analysis of about 18 *D. melanogaster* core promoters, it was observed that mutation of the DPE motif results in upto 50-fold basal transcription activity reduction [75–77].



### 1.5.3 Emerging characteristics of core promoter architecture

Until recently, the core promoter had been presumed to be a generic entity that functions by a single mechanism. However, it is emerging that there is some variation in core promoter structure and function because the TSSs are not located at a single position but are distributed over several to hundreds of bases. This has given rise to two major categories of transcription initiation patterns and thereby core promoters. The first category is the focused/narrow core promoters whose transcription initiates at a single nucleotide or within a region of several nucleotides (Figure 1.4). Narrow core promoters are the predominant mode of transcription in simpler organisms. Secondly, there are broad/dispersed core promoters that harbor multiple weak TSSs over a region of about 50 to 100 nucleotides [42] [44] [72] [78]. Broad core promoters are dominant in vertebrates and they are thought to account for approximately two-thirds of human genes.

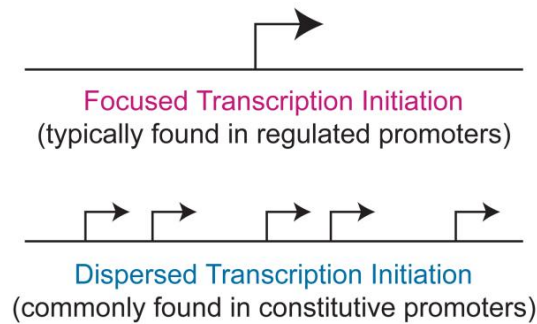


Figure 1.4: Narrow (focused) versus broad (dispersed) core promoters. In narrow core promoters, there is either a single major TSS or several TSS within a narrow region of several nucleotides. In broad core promoters, there are several weak TSS over a broad region [79].

Some broad core promoters exhibit properties of both narrow and broad initiation patterns. For example, a broad core promoter might have multiple TSS with one particularly strong TSS. These are classified as “broad with peak”. “Broad without a peak” exhibit a fairly uniform spread of TSS positions. In vertebrates, narrow core promoters tend to be associated with tissue specific gene promoters. Broad core promoters are typically observed in constitutive (housekeeping) gene promoters, reviewed by Lenhard and colleagues [53].

In *Drosophila melanogaster*, narrow core promoters show propensity for the TATA-box and INR motifs. Broad promoters tend to harbor variably located core promoter motifs. Association of discrete core promoter motifs with either broad or narrow core promoter classes underscores the importance of the motif diversity for transcription regulation [80].



## 1.5.4 Evolution of core promoter architecture

### 1.5.4.1 Comparative aspects of prokaryotic and eukaryotic core promoter organisation

Prokaryotic promoter and transcription regulatory models were initially used as a prototype to describe transcription regulation in eukaryotes. Essentially promoters of prokaryotic organisms such as bacteria and unicellular protists share similar elements to the eukaryotic promoters although there are a few basic differences. They contain at least two conserved features defining the region where the RNA polymerase binds namely; the TSS and the TATA-box. There are some distinguishing features between bacterial and eukaryotic promoters. Firstly, the TATA-box is located at -10 position relative to the TSS in contrast to the -35 position in eukaryotic promoters. Secondly, bacterial promoters harbor the TTGACA sequence, also called the -35 element, located around 35 bp upstream of the TSS. Lastly bacterial promoters also harbor the UP element, located upstream of the 35 element [81–84].

Regarding the mechanistic aspects of transcription regulation, the prokaryotic model is somewhat different to the eukaryotic model in the sense that, a prokaryotic promoter initiates the transcription of several structural genes adjacent to it. This arrangement is referred to as an operon. A single transcribed mRNA is translated into a number of proteins with related functions. In this operon model, promoters have adjacent or interspersed TFBSs to which TFs bind [85] although some cases of transcriptional control using the operon model in eukaryotes have been discovered [86].

As described in section 1.5.3, TSSs in a eukaryotic promoter model are spread out over a longer genomic distance and the number of TFBSs is ordinarily much higher. This may be due to the additional and complex regulatory tasks utilized by multicellular species, for example, maintenance of distinct tissues and cell to cell communication and development. Even though the transcription initiation complexes are similar, key motifs associated with constitutively expressed genes show variations among differ-

ent metazoan groups. The variations in motif usage indicate that at least some of the motifs are derived in a lineage-specific manner and are therefore not an ancient delineator of promoter types reviewed by Lenhard *et al.*, [53].

Plant promoters also share common core promoter motifs with metazoans. The most conserved group is the TATA-box and INR. Using an *in silico* method named LDSS (Local Distribution of Short Sequence), Yamamoto and colleagues [87] identified hundreds of hexamers and octamers with localized distributions within promoters of *Arabidopsis thaliana* and rice. Based on their localization patterns, identified sequences were categorized into three groups namely; pyrimidine patch (Y Patch), TATA box, and REG (Regulatory Element Group). The REG group includes more than 200 sequence motifs half of which corresponded to known cis-elements. In this analysis, the conservation of promoter architecture between monocot and dicot plants was confirmed.

Further analysis of the similarity between plant and mammalian motifs showed that TATA-box sequences are well conserved between plants and mammals but the INR and REG motifs did not exhibit conservation [88]. Despite the difference in the overall INR pattern of mammalian and plants promoters, a dimer consensus at the -1/+1 position, was consistent with the YR Rule, where Y (C or T) at the -1 position and R (A or G) at +1, is applicable to both mammals [71] and plants [87]. Hence, the consensus for TSSs at a dimer level can be considered to be essentially conserved between plants and mammals [88]. In this thesis, the -1/+1 position is interrogated in the context of *G.morsitans*.

#### 1.5.4.2 CpG islands composition

The initiation patterns across genomic windows of varying lengths that has led to the broad/narrow classification (described in section 1.5.3 ) appears to be conserved across the plant and animal kingdoms. Most of the mammalian broad core promoters

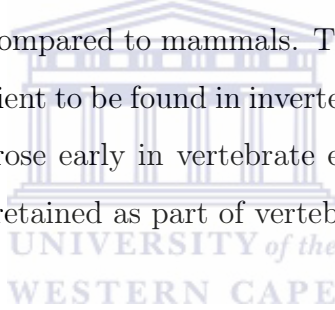
are known to be TATA-less and are associated with CpG-islands [72] [89]. Narrow and broad initiation patterns have also been observed in *D. melanogaster*. According to Rach and colleagues [80], akin to mammals, *Drosophila's* narrow and broad promoters show variations in the enrichment of core promoter motifs. While narrow promoters contain location-specific core promoter motifs such as the TATA-Box, INR, DPE, and MTE, broad promoters were associated with non-location-specific elements as DNA replication element (DRE) and the Ohler1, 6 and 7 [80]. In the same study, mammalian CpG islands are characteristic of broad promoters but they are not found in broad promoters of *Drosophila* spp. This study also established higher frequencies of G and C in *Drosophila's* narrow core promoters relative to broad core promoters suggesting that the functional properties of CpG islands may not dependent on whether a promoter is narrow or broad [80].

Analysis of *Arabidopsis* TSSs on a genome-wide scale showed that the narrow TSSs are conserved between mammals and plants, while broad TSSs with CpG islands in mammalian genomes are found in a plant-specific core promoter type known as the GA type [88] [90]. Though the exact function for CpG islands with regard to transcription is yet to be well defined, they are thought to exert action primarily by destabilizing nucleosomes. This attracts proteins that create a transcriptionally permissive chromatin state. Therefore, CpG islands may 'by design' be equipped to influence local chromatin structure to enable regulation of gene activity [91].

As has been observed with *Drosophila* promoters [92], plant promoters do not harbor CpG islands [88]. However, CpG islands are found in fish genomes. For instance a study by Han and Zhao [93] established variation in both the number and density of CpG islands in four fish genomes (tetraodon, stickleback, medaka, and zebrafish). Accordingly, accumulation of CpG islands around the TSS appears to be a vertebrate specific genomic feature [94]. To establish whether the association of CpG islands to gene promoters is a cause or consequence of evolution, Sharif and colleagues [94] used bioinformatic methods to analyse nine species including invertebrates, chordates

and vertebrates. Their results suggest that CpG islands arose around the TSS as a "consequence" of evolution of the warm-blooded vertebrates, supposedly for efficient regulation of transcription in larger genomes. They further, showed that CpG islands could have functioned as a direct "cause" of evolution among the warm-blooded vertebrates, for example, to facilitate the gain of placenta, which is a unique to some eutherian mammals.

A recent study performed by Okamura and colleagues [95] used *Ciona intestinalis* (an invertebrate that is close to the vertebrates from an evolution perspective) to determine the origins of the CpG-containing vertebrate promoters. A high CpG content around the TSS was observed, but their levels in the promoters and background sequences differed much less compared to mammals. The results suggested that CpG islands are not sufficiently ancient to be found in invertebrates. The authors postulate that CpG islands probably arose early in vertebrate evolution through some active mechanism. They have been retained as part of vertebrate promoters.



#### 1.5.4.3 Effect of TFBSs and TSS turnover

Mutations in protein-coding genes have for a long time remained as the basis for explanation of phenotypic variation between organisms. In 1975, King and Wilson [96] had observed that despite the almost identical sets of proteins, extensive physiological differences could be seen between closely related species. They hypothesized that difference in closely related organisms' physiology may have been due to changes in gene regulation. Several studies have confirmed this hypothesis in various aspects of gene regulation. For example, expression profiles vary greatly even between strains of the same species such as *Drosophila melanogaster* subgroups [97], primates [98] [99], *Arabidopsis thaliana* strains [100] and yeast strains [101]. Evolutionary variations have also been identified in other gene regulatory aspects such as TFBSs in yeast [102,103], humans [104] and vertebrates of different orders [105]. Other gene regulatory aspects with variations include histone modifications [106] and

nucleosome positioning [107–109].

A recent comprehensive comparison of human and amphibian promoter sequences by Van Heeringen and colleagues [110] revealed both similarities and differences in core promoter architecture. Some of the differences originate from a highly divergent nucleotide composition of amphibian and human promoters. Though the distribution of a few core promoter motifs was conserved independently of species-specific nucleotide bias, the recurrence of another class of motifs corresponded with the single nucleotide frequencies. Essentially, this study underscored both the conserved and diverged aspects of vertebrate transcription. Most importantly, it showed preferred motif usage to assemble the transcriptional machinery to promoters with varying nucleotide composition. These observations showed that changes in nucleotide composition exhibit compatibility with conserved transcription initiation mechanisms.

The mechanisms through which regulatory sequences change, yet preserve function continues to be an important open question for evolution [111]. For instance, the even-skipped enhancer system of *Drosophila* species depicts high conservation at the functional level (by, maintaining a high similarity of expression pattern) but substantial divergence at the sequence level. A number of experimental studies suggested that in the even-skipped enhancer region, compensatory mutations are responsible for preserving its functionality in evolution [112–114].

While studying heterotachy (shifts in site-specific evolutionary rates over time) in mammalian promoter evolution, Taylor *et al.*, [115] established that, the degree of promoter evolution differed between lineages and was significantly increased in primates. The study led to the conclusion that the predominant cause is variation in the mutation rate specifically within promoter regions. According to Dermitzakis and Clark [116] and Stone and Wray [117] promoter regions are subject to much less stringent selection and exhibit higher nucleotide substitution rates when compared to protein-coding genes. Consequently, short TFBSs can easily turn over and be

replaced by new ones via random mutations. In most instances however, the functions of the TFBSs may, continue to be well conserved despite substantial sequence changes [118]. In their study of mammalian TFBSs evolution in gene regulatory regions, Dermitzakis and Clark [115] estimated TFBSs turnover rates as high as 32-40% between human and rodent species. Frith and Colleagues [119] established turnover of mammalian TSSs of orthologous genes though at a lower frequency compared to TFBSs turnover. However, a recent study by Main and colleagues [120] evaluated TSS evolution among four *Drosophila* species showing that TSS locations are highly conserved between each species, which is in contrast to mammalian estimates by Frith and colleagues [119].

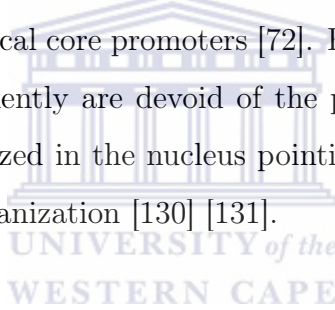
#### 1.5.4.4 Repeat and retrotransposon sequences recruited as promoters

It has become apparent that transcription can originate from regions that have previously been ignored in transcriptome studies. This include a subset of retrotransposons that can act as promoters for tissue-specific non-coding RNAs [121] [122] or as alternative forms of protein-coding mRNAs [123] [124]. According to Jordan *et al.*, [125] approximately 25% of experimentally characterized human promoters contain transposable element insertions including empirically defined cis-regulatory elements. While attempting to define the overall contribution of repetitive elements in mammalian transcription, Faulkner and colleagues [124] estimated that 6-30% of mouse and human RNA transcripts initiate within repetitive elements. These transcripts were found to be generally tissue specific and coincided with gene-dense regions.

Retrotransposon sequences and repeat elements have an effect on promoter evolution as their expansions or contractions can modify the number of and spacing between functional TFBSs [126]. In some studies, it has been argued that larger scale rearrangements such as transposition and duplication can also assemble novel regulatory sequences [43] [127]. The overall effect of such alterations may generate a pool of

phenotypic diversity. Using the *Saccharomyces cerevisiae* model, the mechanisms underlying repeat-based expression divergence have been proposed to originate in chromatin structure. AT-rich promoter repeats were shown to influence local nucleosome positioning. Additionally, variations in the number of repeats were shown to affect the positioning and density of nucleosomes in the crucial part of the promoter [128]. Microsatellite repeats have also been identified in core promoters of plants such as *Arabidopsis thaliana* [129].

Retrotransposons/repeat driven promoters have complex transcriptional regulation as their initiator consensus is different from canonical core promoters. These promoters were found to harbor the consensus (AGT/G) with a complete lack of the initiator dinucleotides found for canonical core promoters [72]. RNAs derived from retrotransposon driven promoters frequently are devoid of the polyadenylation tail. Further, these RNAs are usually localized in the nucleus pointing to a role in transcriptional regulation and/or nuclear organization [130] [131].

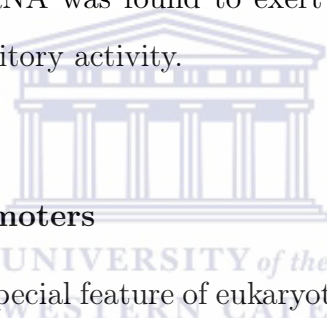


#### **1.5.4.5 Promoter-associated non-coding RNA**

The ENCODE projects results proposed that majority of the human genome is transcribed to generate non-coding RNAs (ncRNAs) [132]. Recent studies suggest that some ncRNAs may exert their function by binding to and regulating the activities of transcriptional co-activator or co-repressor complexes [133]. Non-coding RNAs occurring at or near promoters are broadly classified with respect to their size and location on corresponding TSSs. On the one hand are broad class promoter-associated ncRNAs referred to as promoter-associated long RNAs (PALRs) and promoter-associated short RNAs (PASRs). Their length spans at least 100 nucleotides and they can be transcribed from the same region in both directions [53]. On the other hand is second set of promoter-associated ncRNAs originate at or immediately downstream of the TSS. They are shorter in length compared to PALRs and PASRs and are transcribed in the same direction as their corresponding protein-coding genes. They may also be

transcribed in the reverse direction [53] [134] [135].

It has been shown that the amount of the promoter associated ncRNAs is proportional to the amount of RNA pol II; therefore, their presence is deemed to be a common feature of active promoters [53]. Interestingly, their relative levels exhibit variations upstream and downstream of the TSS and can be linked to different promoter classes. Narrow promoters display a tendency to have more small RNAs downstream of the TSS. Broad promoters display a more even distribution [53] [136]. A long promoter-associated ncRNA has recently been identified to be transcribed from the cyclin D1 promoter upon induction by ionizing-irradiation. The cyclin D1 promoter-associated [137] ncRNA was found to exert transcriptional repression via histone acetyltransferase inhibitory activity.



#### 1.5.4.6 Bidirectional promoters

Bidirectional transcription is special feature of eukaryotic genomes where distance between two TSS of neighboring genes on opposing strands is less than 1 kb [138]. Bidirectional promoters regulate transcription of approximately 11% of human genes [139] and approximately 11% of yeast genes [140]. Bidirectional promoters are also found in plant genomes such as *Arabidopsis thaliana* [141]. The silkworm chorion genes encompass a large group of divergently transcribed gene pairs. Aspects of their bidirectional transcriptional controls have been investigated and they have been proposed as a possible model for future investigation of the intricacies of bidirectional transcription [142]. Studies exploiting comparative genomics to analyse this genomic architecture point toward its conservation in vertebrates and as such, bidirectional promoters may possess distinct biological implication [141], [143–147].

Since bidirectional promoters share common regulatory motifs, co-regulation is believed to be a distinctive feature of bidirectional gene pairs [138]. To elucidate the genomic architecture of bidirectional promoters, Xu and colleagues [148] recently per-



formed a large-scale identification and pathway enrichment analysis of bidirectional gene pairs among several eukaryotes including *D.melanogaster*. Pathway analysis results validated the co-expression of bidirectional genes but did not support their functional relevance. Analysis of the overall evolutionary tendency of bidirectional genomic architecture suggested that the conservation of bidirectional promoters may not be the result of functional bias at whole genome level rather than functional connection between paired genes. By extension, this implies that the genome-wide functional constraint is crucial for the conservation of bidirectional genomic architecture [140].

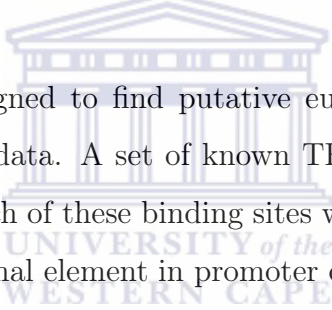
## 1.6 Strategies for TSS identification

### 1.6.1 *In silico* prediction of promoters

The description of a promoter thus far, presents a somewhat linear view of the promoter. In reality, during transcription initiation, a layer of complexity is added by assembling the TFs to adopt a three-dimensional configuration. It is this configuration that facilitates the interaction with other TFs to activate the basal transcription machinery [149]. The main objective of any promoter prediction program is to identify the TSS accurately. Design of such programs is based on the premise that promoter regions have some distinctive features compared to the rest of the genomic sequence. Several machine learning techniques are employed using experimentally validated TSSs as well as presence of specific TFBSs such as core promoter motifs. The program is then used to scan novel genomic sequences while trying to integrate as much information as possible to improve accuracy [150].

These programs include; ProSOM, First Exon Finder (FirstEF), the Neural Network Promoter Prediction (NNPP), PROMOTERSCAN, AUTOGENE, PromFind, CorePromoter, and PromoterInspector. ProSOM is based on unsupervised clustering

of physical properties of DNA [151] while FirstEF employs decision trees and probabilistic models optimized to find potential first donor sites (GT) and CpG-related and non-CpG-related promoter regions based on discriminant analysis. For every potential first donor site and an upstream promoter region, FirstEF determines whether the intermediate region can be a putative first exon, based some quadratic discriminant functions [152]. The NNPP program is a neural network model of the compositional and structural properties of a eukaryotic core promoter. It was developed for analysis of the *D.melanogaster* genome. NNPP's model uses a time-delay architecture, an exclusive case of a feed-forward neural network. This model's structure allows for variable spacing between TFBSs. This variable spacing is recognized as a key player in the transcription initiation process [153].

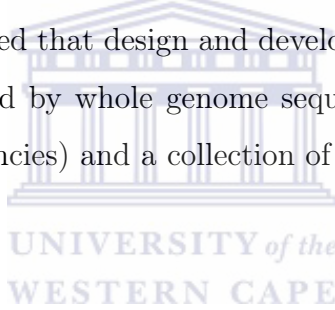


PROMOTERSCAN was designed to find putative eukaryotic RNAP promoter sequences in primary sequence data. A set of known TFBSs was used as background set density. The density of each of these binding sites was then used to derive a ratio of density of each transcriptional element in promoter compared to non-promoter sequences. Individual density ratios of all binding sites were combined and used to build a scoring profile known as the Promoter Recognition Profile. This profile, was used in combination with a weighted matrix for scoring a TATA-box and used to distinguish promoter from non-promoter sequences [154]. AUTOGENE includes a module for promoter identification a method to predict promoter regions in eukaryotic genes through revealing potential TFBSs and analysing patterns for their localisation within promoters [155]. PromFind is not based on any collection of putative TFBSs but, rather, on the variations in nucleotide hexamer frequencies between promoters and protein-coding regions as well as noncoding regions downstream of the first coding exon [156]. Calculation of these variation are based on and a formula first applied by Claverie and Bougueleret [157].

CorePromoter employs a modular approach based on initial localization of a functional promoter into a 1 to 2 kb region from within a large genomic DNA sequence

of 100 kb. From this larger region, the TSS is localised into a 50 to 100 bp (core promoter) region by employing positional dependent 5-tuple measures using a quadratic discriminant analysis (QDA) [158]. PromoterInspector is based on libraries of IUPAC words obtained from training genomic sequences by an unsupervised learning approach. Promoter prediction is based on their genomic context, instead of their exact location [159].

Computational promoter predictors are important for *in silico* gene discovery, but they exhibit high levels of inaccuracy. For instance, a study by Bajic and colleagues [160] demonstrated this inaccuracy on eight promoter prediction programs and concluded that promoter structure variation may not be well covered during algorithm design. They suggested that design and development of promoter prediction programs should be supported by whole genome sequences (due to species specific variation in nucleotide frequencies) and a collection of experimental data with many promoters of different nature.



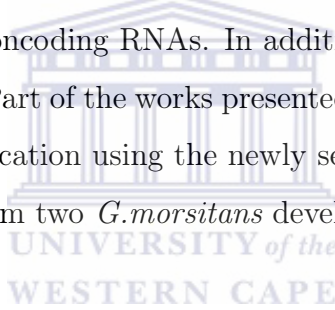
### 1.6.2 Experimental determination of promoters

Conventional experimental TSS identification methods have relied on sequencing of full-length cDNA libraries [161] [162]. These methods can provide evidence on TSS derived from large numbers of 5' end sequences. However, the high costs that would be associated with sequencing a comprehensive cDNA library is a limitation to the throughput that a cDNA library would demand to provide significantly sufficient data for lowly expressed genes.

High-throughput next generation sequencing (NGS) technologies have been employed to facilitate the elucidation of transcriptional control mechanisms through promoter identification and expression profiling. One such method is cap analysis of gene expression (CAGE) which allows high-throughput identification of sequence reads corresponding to the 5' end of mRNA at the cap site and the identification of tran-

scriptional start points [163]. The method employs the Cap trapper full-length cDNA technology [164] followed by cleavage of the first 20 base pairs, PCR, concatemerisation, cloning and ultimately, sequencing of the CAGE reads/tags. TSS-seq [165] is a recently developed method that employs NGS technologies and 5' capping. TSS-seq employs the oligo-capping full-length cDNA technology [166] while CAGE employs the Cap trapper full-length cDNAs technology [164]. In addition, the read lengths attainable by TSS seq are slightly longer (36 nucleotides) compared to CAGE (20-21 nucleotides).

NGS-based TSS identification methods provide accurate high-throughput measurement of RNA expression by allowing mapping of all the transcription initiation sites of both capped coding and noncoding RNAs. In addition, TSSs are characterized at single-nucleotide resolution. Part of the works presented in this thesis aim at developing a methodology for TSS location using the newly sequenced *G.morsitans* genome and TSS-seq tags sampled from two *G.morsitans* developmental stages.



## 1.7 Strategies for identification of transcription factor binding sites

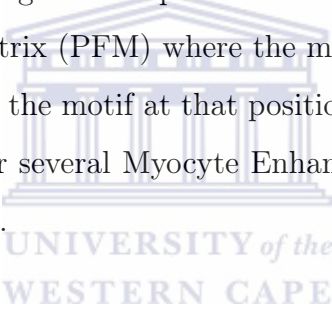
### 1.7.1 Computational prediction of TFBSs

Accurate TSS identification is followed by promoter extraction after which promoter motifs/TFBSs are characterized as the first step towards deciphering the DNA regulatory code. Even though novel approaches are continuously proposed and have reported some successes in lower organisms, the problem of computational prediction of TFBSs still remains a major challenge for higher organisms. This is because of an intricate organization and degeneracy of the genetic code. TFBSs are very short, and can tolerate high degrees of degeneracy in the sequence, resulting in high rate of false positive prediction. In addition, albeit continued refinements on prediction of

TFBSs, the cellular environment modulates transcriptional events by imposing the chromatin structures' selective constraint [167].

### 1.7.1.1 The position weight matrix

The position weight matrix is arguably the most commonly used representation of motifs in biological sequences. Development of a PWM stems from simple logic that asks the question, “Given a list of TFBSs for a specific TF, how would one best represent and describe the information contained in these sites for further analysis?” Eventually, one would aim at finding a representation that matches all the possible TFBSs in the list that is distinct from a background sequence. A consensus sequence is generated using a position frequency matrix (PFM) where the most frequent character at each position is chosen to represent the motif at that position. For example, the following is an illustration of a PFM for several Myocyte Enhancer Factor-2 (MEF2) TFBSs; (Example adopted from [168]).



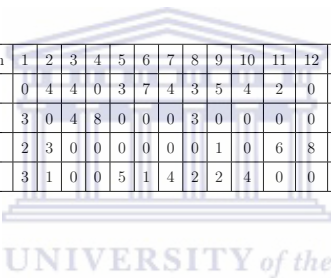
position	1	2	3	4	5	6	7	8	9	10	11	12	13	14
site1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
site2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
site3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
site4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
site5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
site6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
site7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
site8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
consensus	B	R	M	C	W	A	W	H	R	W	G	G	B	M

Figure 1.5: PFM for several Myocyte Enhancer Factor-2 (MEF2) TFBSs [168].

The information contained in the various cells in the matrix shows that some positions may have equivalent frequencies thus the IUPAC sequence [169] is used. This representation is suitable for short and/or highly conserved motifs, because albeit arbitrarily, it is well defined and contains much of the information from the original

binding site. However, while working with longer and/or degenerate motifs the consensus sequence method may not suffice. It may lead to failed recognition of *bona fide* TFBSs and/or a relatively high rate of spurious matches.

Position weight matrices (PWM) (also known as position specific scoring matrices (PSSMs)) provide quantifiable descriptions of the known binding sites for a TF and hence a more precise representation [170]. Construction of a PWM begins with the PFM where the width of a PFM is equal to the length of the sequence. Assuming all the sites have the same length in general (which is not true) The PFM for the MEF2 TFBS above would be presented as;



position	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
G	3	0	4	8	0	0	3	0	0	0	0	0	2	4
C	2	3	0	0	0	0	0	1	0	6	8	5	0	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

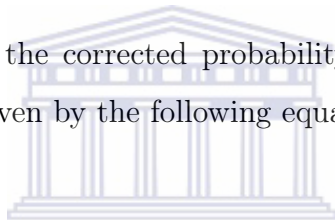
Each coefficient in the matrix is an indication of the number of times that a specific nucleotide has been observed in a given position. A PWM is constructed by calculating the probabilities of observing each nucleotide at each position. Thus for a given motif of length  $l$ , the product of the coefficients from such a matrix corresponding to each nucleotide at a given position of the sequence will be the probability of finding such a motif in a true functional site.

In an ideal scenario one would expect that each of the nucleotides is equally probable (in which case the matrix would have the probability  $0.25^9$ ). However, since nucleotide distribution varies between genomes the concept of likelihood ratio is employed where the nucleotide distribution in a given genome becomes the expected background probabilities. The probability of finding a motif in a random site is the product of the *a priori* probabilities of the corresponding nucleotides. The likelihood ratio is the ratio between the probability of a sequence in a functional site and the probability of the sequence in a random site (if nucleotide distribution was equiprob-

able, such a ratio would be zero).

For efficient computation, the probabilities are converted to a log scale. The log likelihood ratio is equal to zero if a motif has the same likelihood of appearing in a functional site than in an arbitrary site. Further, the log likelihood ratio is smaller than zero if the motif is more likely to be found in an arbitrary site than in a functional site. To eliminate null values before log likelihood ratio calculation, and somewhat correct for small samples of binding sites, a sampling adjustment, known as pseudocounts is added to each cell. There is no definite formula for calculating pseudocounts and this varies from software to software [171].

By applying the above rules the corrected probability of observing a specific motif given a normalized PFM is given by the following equation;



$$p(b, i) = \frac{f(b, i) + s(b)}{N + \sum_{b' \in A, G, C, T} s(b')} \quad \text{equation(1.1)}$$

where  $b' \in A, G, C, T$

$p(b, i)$  = corrected probability of base  $b$  in position  $i$

$f(b, i)$  = counts if base  $b$  in position  $i$

$N$  = number of sites

$s(b)$  = pseudocount function

Dividing the nucleotide probabilities in equation 1.1 by the expected background probabilities and converting the values to a log scale, the PWM calculation for each element in the matrix is presented by this equation;

$$w(b, i) = \log_2 \frac{p(b, i)}{p(b)} \quad \text{equation(1.2)}$$

$p(b)$  = background probability of base  $b$

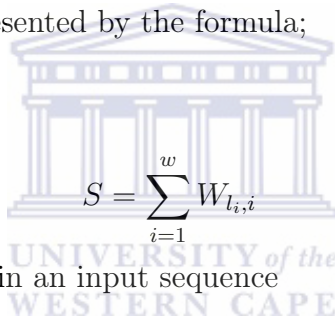
$p(b, i)$  = corrected probability of base  $b$  in position  $i$

W  $b,i$  = PWM value of base  $b$  in position  $i$

By substituting the PFM values using this equation 1.2 the MEF2 PFM is transformed as shown below;

position	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	-1.93	0.79	0.79	-1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	-1.93	-1.93	0.79
G	0.45	-1.93	0.79	1.68	-1.93	-1.93	-1.93	0.45	-1.93	-1.93	-1.93	-1.93	0.00	0.79
C	0.00	0.45	-1.93	-1.93	-1.93	-1.93	-1.93	-1.93	0.66	1.30	1.30	1.68	1.07	-1.93
T	0.15	0.66	-1.93	1.93	1.07	0.66	0.79	0.00	0.00	-1.93	-1.93	-1.93	-0.66	-1.93

A quantitative score for a potential motif is produced by summing the values for each cell in the matrix and is represented by the formula;



$$S = \sum_{i=1}^w W_{l_i,i}$$

equation(1.3)

$l_i$  = the nucleotide position  $i$  in an input sequence

$S$  = PWM score of a sequence

$W$  = width of the PWM

Every potential binding site in both the positive and negative orientation of the DNA strand is evaluated by sliding the PWM over the sequence in one base pair increments for longer sequences. A threshold is specified to assess whether the input sequence matches the motif or not. The score for the MEF2 motif is shown below;

Site scoring	0.45	-0.66	0.79	1.68	0.45	-0.66	0.79	0.45	-0.66	0.79	0.00	1.68	-0.66	0.79
Consensus	T	T	A	C	A	T	A	A	G	T	A	G	T	C

The information content of a row or column is the sum of information contents of its corresponding cells and it is used to assess specificity of each column of the alignment and it represents the capability of the matrix to distinguish between a binding site



(the matrix itself) and the background model. The information content for every cell in the matrix is calculated by multiplying the weight by the frequency with the equation below;

$$D(i) = 2 + \sum_b P_{bi} \log_2 P_{bi} \quad \text{equation(1.4)}$$

$D_i$  = information content in position  $i$

$P(b,i)$  = corrected probability of base  $b$  in position  $i$

Most of the popular methods that use PWMs to model DNA motifs use the information content measure to identify the optimal motifs from input sequences [172–175]. Information content is presented graphically by sequence logos to facilitate fast and intuitive visual assessment of pattern [176]. Basically a sequence logo scales each nucleotide by the cumulative bits of information multiplied by the proportionate occurrence of the nucleotide at a given position.

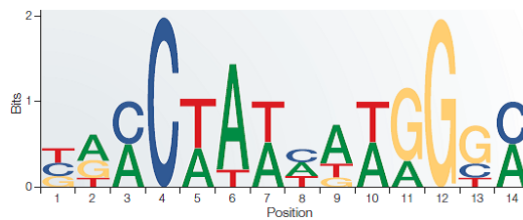
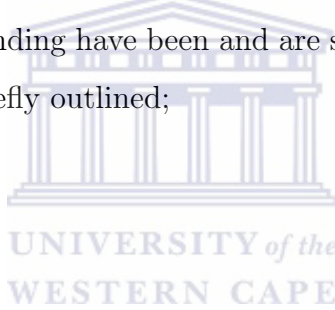


Figure 1.6: Sequence logo depicting the information content of MEF2.

The main weakness of the PWM model is the assumption that nucleotides at each position are mutually exclusive. In effect, the fitness score of a matched motif with its profile is the sum of fitness at each position. However, in some cases, position dependence exists on the TFBSs and this has been verified experimentally and/or statistically. For example, when Ellrott and colleagues [177] applied chi-squared test on the 71 binding sites of TF hepatocyte nuclear factor 4 (HNF4), a significant dependence was found between several pairs of positions including positions 4 and 8,

4 and 11. Additionally, it has been shown that nucleotides for the TFBS motif of TF Early growth response protein 1 (Egr1) cannot be treated independently [178]. Several other representations of binding sites that encapsulate position dependence among nucleotide positions have been proposed [179–181]. Secondly, the PWM model assumes that TFs have strict spatial requirements in their binding sites that prevent variable spacing. TFs, such as a subset of the nuclear receptor family exhibit variable spacing rendering standard PWMs inappropriate for TFBS prediction [182]. Specialized models, such as one for the TF CCAAT box-binding transcription factor (CTF) have been created to represent binding for some of these cases [183].

More comprehensive representations of PWMs that encompass the potential dependence between positions in binding have been and are still under development. Their underlying algorithms are briefly outlined;



#### 1.7.1.1.1 Markov chains

A Markov chain is a sequence of arbitrary values whose probabilities at a time interval relies upon the value of the number at the previous time. It can be presented by the figure below;

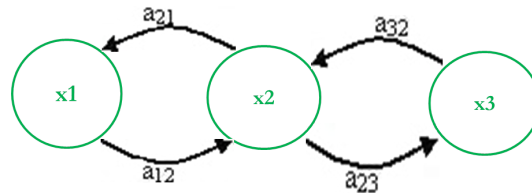


Figure 1.7: Representation of the markov chain concept.  $x_1$ ,  $x_2$ ,  $x_3$  represent states while  $a_{12}$ ,  $a_{21}$ ,  $a_{23}$ ,  $a_{32}$  represent transition probabilities.

A markov chain of length  $n$  describes the probability distribution of  $n$  states  $x_1, x_2, \dots, x_n$  by means of transition probabilities where the transition probability of a sequence is

simply the probability of going from one state to another. Transition probabilities may take many paths for instance, state  $x_1$  to state  $x_3$  or back to itself. Ultimately, the sum of the transition probabilities must always equal to 1. With regard to DNA, the states ( $x$ ) would present any of the four nucleotides ACGT whilst the transition probabilities would be, the probability of one nucleotide following another nucleotide. The main property of a markov chain is that the probability of each symbol  $x_i$  depends only on the value of the preceding symbol  $x_{i-1}$ , not on the entire previous sequence [184].

A Hidden markov model (HMM) is a statistical model where the system being modeled is considered to be a Markov process with unknown parameters or unobserved states. The figure below presents a hidden markov concept where;

$x_1, x_2, x_3$  = hidden/unknown states

$y_1, y_2, y_3$  = observable states

$a_{12}, a_{21}, a_{23}, a_{32}$  = transition probabilities

$e_1, e_2, e_3$  = emission probabilities

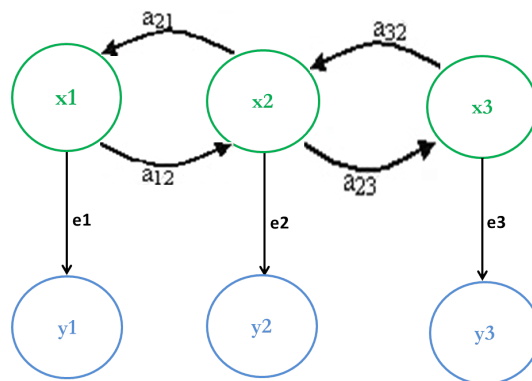
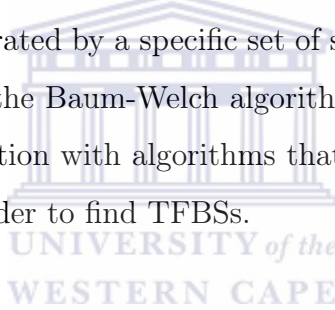


Figure 1.8: Representation of the hidden markov concept.

Green circles represent hidden states while the blue circles represent the observable states.

Importantly, only the blue circles are seen but they directly are dependent on some hidden state (green circles). The challenge usually is to find out the hidden states, the emission probabilities and transition probabilities. In a genomic context, a HMM considers the probability of a nucleotide occurring in all the different states possible in a sequence. Each state contains its own set of probabilities as it assumes the function of a sequence determines its composition, in that sense a HMM determines the probability that a sequence has its order and composition of nucleotides [184].

HMMs are used in modeling background genomic sequences that are crucial to enhance predictive accuracy of any motif finder. The background genomic sequence is modeled by one set of states and binding sites by a different set. On the positive strand, binding sites are generated by a specific set of states to those on the negative strand usually by employing the Baum-Welch algorithm [185]. Motif finders use the background model in conjunction with algorithms that consider the logical and spatial relationships of TFs in order to find TFBSs.



The basic markov chain is the first order markov chain, which estimates the probability of finding a given nucleotide based on the previous nucleotide. In effect, a second order markov chain will determine the probability of finding a given nucleotide based on the previous two nucleotides. This can be generalized as “an  $n$ th-order markov chain model determines the probability of a nucleotide based on the last  $n$  nucleotides” [184]. Higher ordered markov models have been shown to improve the discriminative power [186]. Third order markov models are especially useful in scanning genomes for motifs as they are capable of scanning 4 letter words that may be of functional significance [187]. The transition probabilities vary from one organism to another since genome nucleotide composition varies thus markov models show strong variations between organisms [188].

### 1.7.1.1.2 Expectation maximization (EM)

This method has also been used to optimize PWMs. It is used in a set of unaligned DNA sequences where each sequence must contain at least one common site and no alignment sites are required. Instead the ambiguity in the location of the sites is handled by the missing information principle to develop an “expectation maximization” [173].

The EM algorithm begins with a speculation of the PWM/motif, which could be entirely random or based on some previous knowledge about the binding sites. For instance, the MEME program uses various sequence models make assumptions about how and where motif occurrences appear in the dataset. The OOPS model is the simplest one as it assumes that there is exactly one occurrence per sequence of the motif in the dataset. The sequence model consists of two components that model the motif and the background respectively using discrete random variables. The background positions in the sequences are modeled by one discrete arbitrary variable. Using the PWM/motif, the probability that each subsequence is a binding site is assessed and the PWM/motif is then redetermined based on the site probabilities. EM estimates the number of occurrences of a given motif in each sequence in the dataset and outputs an alignment of the motif instances. In this way, EM is capable of discovering several different motifs with different instances in a single dataset [189].

The main strength of the EM algorithm is the fact that it has flexible options and matrices are scored with E value. The lower the E value the higher the quality of the match. EM supports multiple matrix widths and higher order markov chains returning the most informative. However, computing time increases quadratically with the size of the sequence set [190].

### 1.7.1.1.3 Gibbs sampling

Gibbs sampling is a stochastic adaptation of the EM algorithm [175] that employs two data structures, one that holds the background sequence and the other that holds the motif in question. One data structure holds the pattern that is represented in the form of a probabilistic model of residue frequencies for each position. This pattern is accompanied by an equivalent probabilistic description of the background frequencies with which residues occur in sites not described by the pattern. The other data structure comprises of the alignment and is a set of positions for the common pattern within the sequences. The goal is to identify the most well-defined and the most probable common pattern that is obtained by locating the alignment that maximizes the ratio of the matching pattern probability to background probability [175].

The algorithm is initialized by selecting random starting positions within the candidate sequences and then proceeds through several iterations to execute a predictive update and a sampling step. In the predictive update step, one of the sequences say  $z$  is selected arbitrarily or in a specified order. The pattern description and background frequencies are then calculated from the all current positions in all sequences excluding  $z$ . In the sampling/stochastic expectation step, all possible segments of a given width within the sequence  $z$  are considered as possible occurrences of the pattern. The probabilities of generating each segment according to the current motif probabilities are calculated as the probabilities of generating these segments by the background probabilities. Ideally, the more accurate the predictive step is, the more accurate the determination of its location during the sampling step and vice versa [191].

Gibbs sampling tends to give a more robust optimization of the PWM as its stochastic procedure enables the minimization of false positives [192]. Gibbs sampling is fast and gives probabilistic description of the patterns. However due to its stochastic nature it often returns a different result on each run. In addition the algorithm predicts frequent false positives, as there is no threshold on pattern significance [190].

Due to its intuitive representation and fast computation the basic PWM is still the leading model in the search for discovering potential TFBSs [193]. Indeed, it has been shown to be equivalent at least, and in some case outperforms, other more complicated models [194]. Various implementations of the aforementioned algorithms alongside their accuracy in various settings are described by Tompa and colleagues [195].

### 1.7.2 Experimental determination of TFBSs

Confirmation of TFBSs via biochemical experimental assays remains the highest form of validation because it provides precise information about the inferred biological function of a TFBS as well as its corresponding TF [167]. Most experimental approaches depend on computational frameworks to detect binding sites and are classified under two systems namely; *in vivo*-based and *in vitro*-based methods. *In vitro* methods identify TFBSs together with the binding energy landscapes. They include assays such as: SELEX [196] [197], DNA immunoprecipitation (DIP-chip) [198] and the classical gel shift assays such as DNAase footprinting [199] and Electrophoretic mobility shift assays (EMSA) [200]. *In vivo-based* methods identify information on the TF consensus binding sites together with the biological context of DNA specific interactions. They include chromatin immunoprecipitation (ChIP) [201] combined with one of ChIP-chip [202] or ChIP-seq [203].

Experimental determination has its limitations. For instance, EMSA does not address the question of which TFBSs in the genome is biologically functional. In addition, while ChIP assays provide a much improved assessment of the TFBS regulatory potential of a given TF, they have high rates of false positive prediction. This is in part because of binding to other chromatin components [204].

Limitations of computational and experimental methods of promoter and TFBSs identification methods dictate that successful investigation of genomic DNA regula-

tory potential requires a cross-disciplinary approach, essentially collaboration between wet-lab and computational scientists (Figure 1.9). Such efforts would yield better results by producing experimental data used to model computational tools. Constant cycles of these collaborations would assist in more accurate computational models.

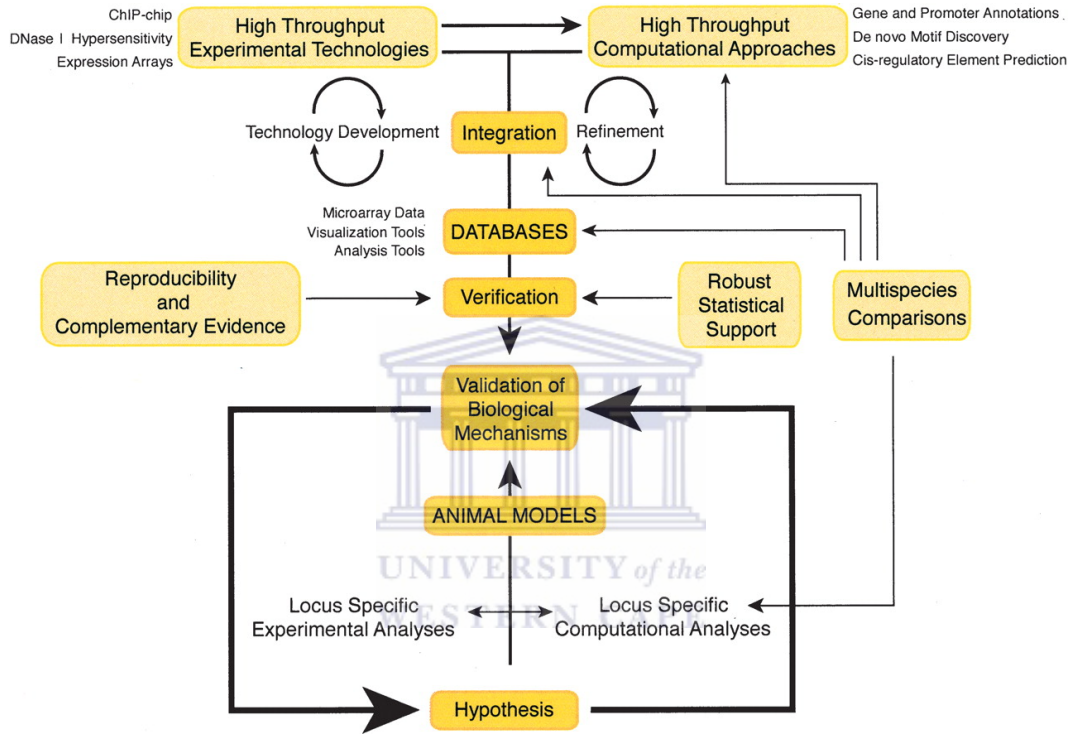


Figure 1.9: The interplay between computational and experimental techniques. During elucidation of genomic transcriptional control elements, results from high throughput experimental and computational approaches are integrated and compiled into databases after which validations are performed and the cycle is repeated [167].

Information accrued from these analyses is stored in databases that are subsequently utilized for follow-up and/or new studies. Experimentally verified binding sites are compiled in databases such as TRANSFAC [205], JASPAR [206], ORegAnno [207] and PAZAR [208]. Databases such as TRED [209], TRRD [210], ooTFD [211] and MPromDb [212] are maintained to facilitate identification of regulatory signals in newly sequenced genomes, each of them focusing on distinct aspects of transcriptional control and the degree to which the sites are studied experimentally. Organism-



specific data detailing TFBSs profiles accrued from experiments can be incorporated as regulatory tracks in various genome browsers.

## 1.8 The *Drosophila melanogaster* model for transcription regulation in insects

*D.melanogaster*, with its well-annotated genome and long history of experimental studies is one of the best characterized dipteran genomes in terms of functionally annotated regulatory elements. Indeed, there have been numerous experimental gene-by-gene studies on promoter elements as well as studies on a genome-wide scale. Following release3 of the *D. melanogaster* genome, Ohler and colleagues [213] identified several known core promoter elements while Gershenzon and colleagues [214] used statistical analyses to describe several motifs with the features of promoter elements including possible novel core promoter elements. Using cap-trapped expressed sequence tags, Hoskins and colleagues [92] produced a high-resolution map of promoters in the *D. melanogaster* genome revealing that TSS distributions of alternative promoters form a complex range of shapes where discrete promoter shapes constitute given TFBSs combinations.

The ModENCODE project comprehensively mapped transcripts, histone modifications, chromosomal proteins, TFs, replication proteins and intermediates across multiple cell lines in various stages of development in *D. melanogaster* leading to the generation of massive datasets that more than tripled the annotated portion of the *D. melanogaster* genome [215]. The interrelated activity patterns of these elements showed stage and tissue-specific regulators. It is anticipated that the datasets generated as part of the ModENCODE project will facilitate directed experimental and computational studies on *D. melanogaster* and related species towards comprehensive genomic and functional annotation. For example, some studies have already utilized

these datasets to study gene regulation including an analysis of chromatin landscape of heat shock factor binding [216], chromatin signatures of *D.melanogaster's* replication program [217], microRNAs [218] and regulation of *D. melanogaster's* glutamate receptor [219].

Perhaps the hallmark of the modENCODE project was the comprehensive annotation of the cis-regulatory map of the *D. melanogaster* genome [220]. The study utilized chromatin immunoprecipitation and histone deacetylases datasets across a developmental time course to infer more than 20,000 candidate regulatory elements. Additionally, this study validated a section of predictions for enhancers, insulators and promoters *in vivo* by identifying approximately 2000 genomic regions of dense TFBSs associated with chromatin activity and accessibility. The study also discovered numerous new TF co-binding relationships and TF networks.

The modMine database [221] has been built by the modENCODE data coordination to facilitate searching and download of datasets to proceed with fine-grained analysis [222].

## 1.9 Thesis rationale and objectives

Sustainable control of HAT would utilise an integrated approach that exploits the transmission cycle on two fronts: human-Trypanosome and Tsetse-Trypanosome interactions. One of the leading difficulties in anti-Trypanosome drug development had been the inability to identify suitable targets from a small percentage of the whole complement of known Trypanosome genes. However, Trypanosome genome sequencing and analyses aided the development of several new approaches for drug target identification. Post genome follow-up studies are continuously garnering comprehensive and accurate target identification using an array of *in-silico* and experimental approaches.

To leverage such achievements on the vector front, the *Glossina morsitans* genome has been sequenced. It is hoped that exploiting genomic information to understand Tsetse biology will offer new and efficient approaches for vector control. For instance, the genome data of *Glossina morsitans* may help unravel some of the underlying mechanisms that are crucial for vector-parasite interactions such as immune responses implicated in parasite elimination. Other mechanisms crucial for vector-parasite interactions include those facilitating efficient blood feeding and parasite maturation in the salivary gland.

The aforementioned processes are facilitated by a series of pathways and interactions whose activation in a time and tissue specific manner is fine-tuned by regulatory mechanisms. Regulatory mechanisms operating at transcriptional level are arguably the most critical as they exert fundamental control over the abundance of nearly all functional macromolecules for any given cell. The core promoter is the minimal promoter fraction that is required for correct assembly of the transcription initiation machinery. On the other hand, the proximal promoter constitutes almost all of functional transcription factor binding sites for any given gene. The immune response adaptations of blood feeding insects to eliminate parasites are reflected in the adap-

tations present in the regulatory regions of immune genes. Testing this hypothesis requires an accurate delineation of the transcription start sites for *Glossina* genes and specifically immune response genes.

The broad aim of the work presented herein was to refine the transcriptional regulatory regions of newly annotated *Glossina morsitans* genome in order to gain insights into the mechanisms of transcription regulation *Glossina morsitans*. To help tackle this, two datasets incorporating transcription start site information from our collaborators (Serap Aksoy-Yale School of public health and and Yutaka Suzuki- Department of Medical Genome Sciences, University of Tokyo) were utilized. These datasets allowed the characterization of *Glossina morsitans* promoter architecture with a view to understanding the basal transcription initiation programs as well as transcriptional control programs of immune genes. The questions poised are presented as distinct studies that can be summarised by the following objectives:

- 1) To utilize *in-silico* approaches to develop a method of TSS identification using experimental data derived using the TSS-seq method.
- 2) To investigate *Glossina morsitans* basal transcription machinery by comprehensive analyses of core promoter properties.
- 3) To computationally identify regulatory motifs in the proximal promoters of *Glossina morsitans* immunity genes.
- 4) To create a data repository for *Glossina morsitans* promoter sequences.

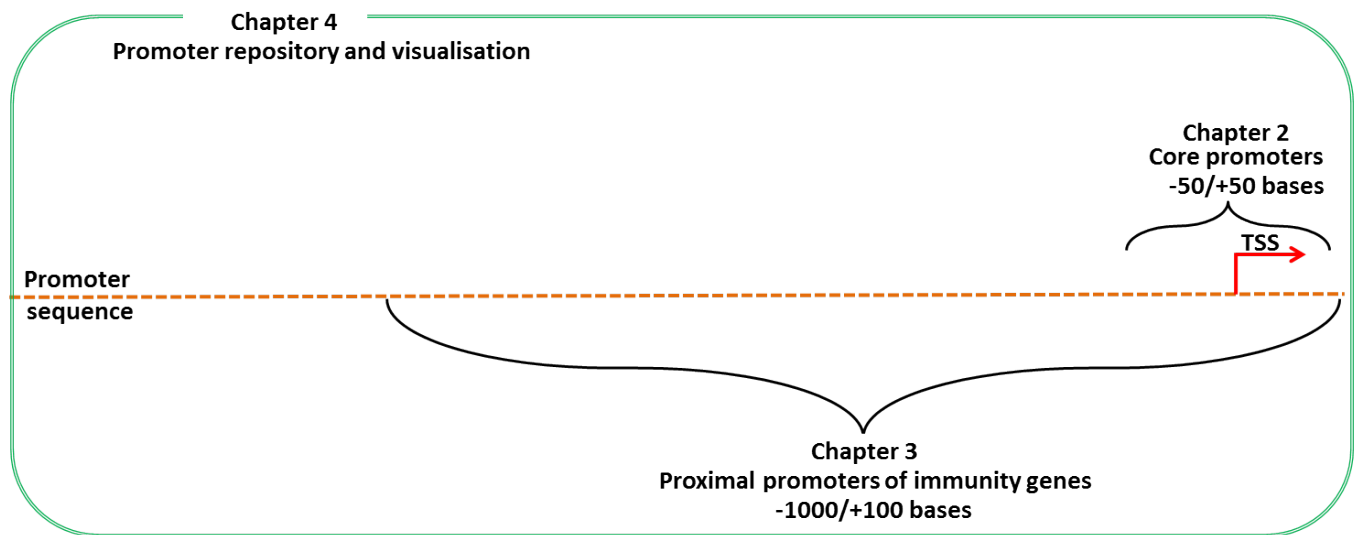
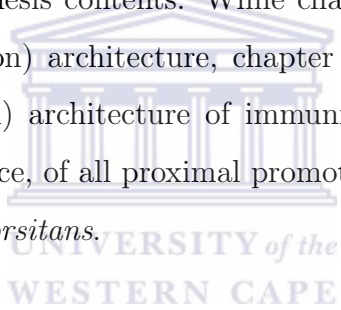


Figure 1.10: Illustration of thesis contents. While chapter two examines the global core promoter (-50/+50 region) architecture, chapter three examines the proximal promoter (-100/+1000 region) architecture of immunity genes. Chapter four is a repository, with a user interface, of all proximal promoters identified by this analysis in the genome of *Glossina morsitans*.



## Chapter 2

# Analysis of core promoter motifs in the genome of *Glossina morsitans* using TSS-seq.



### Abstract

**Background:** The principal mechanism of transcriptional control occurs at the promoter level through the interaction of sequence-specific interactions of DNA-binding proteins known as transcription factors with cis-elements. Canonical core promoter motifs required for basal transcription initiation consist of the BRE, TATA-box, INR, DPE and MTE motifs. Delineating these motifs require accurate identification of the transcription start site. High-throughput methods for transcription start site detection on a genome-wide scale have shown that eukaryotic transcription initiates across genomic windows of varying lengths. Thereby, core promoters have been classified as narrow or broad based on their respective genomic distance. These core promoter categories exhibit variation in terms of motif composition as various motifs exert their functions in a combinatorial mode and exhibit preference across various promoter classes.

**Methods:** In this study, a method for TSS identification in the newly sequenced *Glossina morsitans* genome was developed using TSS-seq tags sampled from two

developmental stages of *Glossina morsitans*. TSS-seq tag clusters mapping to the predicted 5'UTR and the first coding exons were used to define transcription start sites and extract core promoter regions. The core promoter regions were classified as narrow or broad based on the number of TSS positions within a TSS-seq cluster and analysed for variation in canonical core promoter motif profiles.

**Results:** A total of 3134 tag clusters were obtained where each cluster was defined with a minimum of 100 reads per cluster. Approximately 45.4% (1424) of the tag clusters mapped to the first coding exons or their proximal predicted 5'UTR regions and include 31 tag clusters that mapped to transposons. A total of 1101 (35.1%) mapped outside the candidate genic regions and/or scaffolds without gene predictions. These tag clusters may correspond to previously un-annotated transcripts or noncoding RNA TSS. The remainder (609 tag clusters) mapped to other genic regions (introns, exons and 3'UTR).

After excluding transposon derived core promoters, 1393 core promoters remained and were used for subsequent analysis. These core promoters exhibited propensity for the AT dinucleotides in contrast to mammalian promoters that exhibit a propensity for CG dinucleotides. Majority (95%) of the core promoters analysed in this study were of the broad type while only 5% were of the narrow type. Comparison of core promoter motif occurrences between random and *bona fide* core promoters showed that, generally, the number of motifs in biologically functional genomic windows in the true dataset exceeded those in the random dataset ( $p \leq 0.00164, 0.00135, 0.00185$  for the narrow, broad with peak and broad without peak categories respectively).

Narrow core promoters recorded higher frequency of the TATA-box and INR motifs and two-way motif co-occurrence showed that the TATA-box-INR pair is over-represented in the narrow category. Broad core promoters showed higher frequency of the BREd and MTE motifs and two-way motif co-occurrence showed that MTE-DPE pair is over-represented in broad core promoters. TATA-less promoters account

for 77% of the core promoters in this analysis. TATA-less core promoters showed a higher frequency of the MTE and INR motifs in contrast to observations in *Drosophila* where the DPE motif has been reported to occur frequently in TATA-less promoters. These motif combinations suggest their equal importance to transcription in their corresponding promoter classes in *Glossina morsitans*.

Functional analysis identified ontologies associated with developmental functions and showed that genes associated with development are controlled by both narrow and broad transcription initiation mechanisms.

**Conclusions:** This work has identified 87 transcription modules not previously identified in *Glossina morsitans* scaffolds as evidenced by the absence of gene predictions on these scaffolds. A total of 1393 experimentally confirmed TSS locations were identified on ENSEMBL predicted genes. These TSS' are being integrated with the VectorBase genome browser for *Glossina morsitans* ([www.vectorbase.org](http://www.vectorbase.org)). The data also demonstrates the occurrence of transposon-mediated promoters of at least 31 cases. The study has established variation in frequency of motif co-occurrence across various initiation patterns. In addition, the variation in dinucleotide composition suggests a fundamental difference in global promoter architecture between mammals and insects. The results elucidate the compositional properties of core promoters in *Glossina morsitans* as a prerequisite to understanding how the RNA polymerase II transcription program is coordinated in *Glossina morsitans*. Sequence data and results presented herein will facilitate analysis of transcriptional control programs in the yet to be sequenced *Glossina* genomes.



## 2.1 TSS-seq and promoter identification

A transcriptional start region (TSR) can be defined as a region in the genome where experimental evidence for transcription initiation exists. Experimental information about TSRs is derived via individual full-length cDNAs or high throughput NGS methods such as CAGE [163] or the recently developed TSS-seq [165]. TSR elucidation methods utilize oligo-capping, a process that entails substituting the cap structure unique to the 5' end of eukaryotic mRNA with an artificial oligonucleotide in a sequence of elaborate steps outlined in Figure 2.1 below. The end product of the oligo-capping procedure is a library specifically enriched for full-length cDNAs corresponding to the region from the 5' end through to the 3' end of the full-length mRNAs.

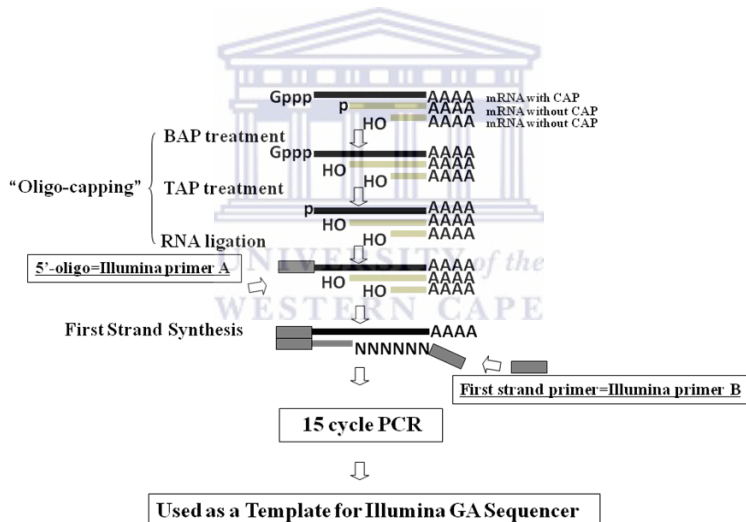


Figure 2.1: Outline of the TSS-seq procedure. Oligo-capped mRNA's are used to produce a library specifically enriched for full-length cDNAs which are amplified via PCR and directly sequenced without cloning. Gppp = cap structure, AAA = poly A tail, BAP=Bacterial alkaline phosphatase, TAP= Tobacco acid phosphatase [223].

The TSS-seq technique was developed by combining oligo-capping and Illumina GA sequencing technology [224]. In this method, primer sequence necessary for the sequencing is directly introduced at the 5' end of capped transcripts by replacing the cap structure with a cap-replacing oligo as shown in Figure 2.1. cDNA is synthesized using random hexamers, amplified with 15 cycles of PCR and directly introduced into

the sequencer without cloning. The output consists of 36 base long reads generated by the sequencer at the rate of 10-30 million reads per run [165].

Given that each transcript has only one cap structure, transcript numbers generated by TSS-seq are proportional to gene expression levels (digital expression profiling). Accordingly, besides locating TSS, TSS-seq enables analysis of the expression levels of transcripts in an extremely high-throughput fashion. The output of a TSS-seq experiment is a collection of reads exclusively enriched with 5' end of transcripts thus mapping TSS-seq reads on a genome appears as a peak signal. The peak position unequivocally directs to where each transcript starts. Due to its high throughput nature, TSS-seq enables clear analysis of variation at a TSS when compared to a full-length cDNA whose read number may ordinarily not suffice for statistical evaluation. TSS-seq has been employed in various studies, for example: to understand transcriptional landscapes [165], to facilitate promoter location and thereby description of transcription regulatory motifs [110], to demonstrate that integrative interpretation of transcriptome data is necessary for the identification of putative promoters [416].

## 2.2 Properties of eukaryotic core promoters

Application of NGS technologies such as CAGE and TSS-seq revealed that most eukaryotic genes do not conform to the simple model in which a TATA-box directs transcription from a single defined nucleotide position. Instead, most genes have multiple TSS and by extension multiple core promoters. Subsequently, core promoters are now described by their start site usage distribution. The distribution may be narrow in the case where transcription initiation proceeds from a single nucleotide or within a region of several nucleotides. On the other hand the distribution may be broad where transcription initiates in a region of 100-200 bases [53]. It is important to note that broad promoters are conceptually distinct from alternative promoters where core promoters are separated by larger genomic space (usually not less than

100 nucleotides).

Generally, narrow core promoters are observed in genes expressed in a tissue-specific expression mode. Most narrow promoters harbor the TATA-box and INR motifs. Broad promoters are associated with constitutively expressed genes and tend to harbor variably located core promoter motifs [80] [226]. It is assumed that TSS positions in broad promoters are defined in part by the positions of the core promoter motifs.

The universal mechanism of basal transcription initiation begins by binding of TATA binding protein (TBP) to the TATA-box. In the case where the TATA-box does not exist, (TATA-less core promoters), TATA associated factors (TAFs) bind to any other core promoter motif present at a functional position and/or to other TFs in order to involve TBP in pre-initiation complex [76] [227–229]. *In vitro* studies have shown that the the TATA-box has the ability to direct transcription initiation alone. However, the rest of the core elements usually work in cooperation with others. Indeed, two-way DPE-INR, MTE-INR, MTE-DPE, BRE-TATA, and INR-TATA synergism have been experimentally established [69] [76] [229–233].

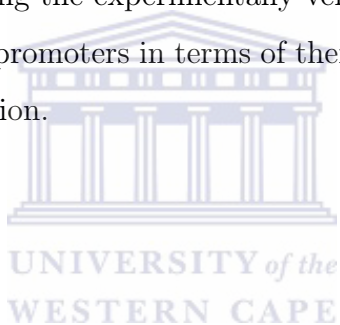
Only a few promoters have been experimentally examined, even for the well-studied core promoter motifs such as TATA-box, INR and DPE. This is predominantly because of the laborious and time-consuming nature of experimental analysis of promoter elements [214]. Statistical approaches have been valuable in complementing experimental analyses by identifying over-represented motifs. For example, Ohler and colleagues [226] [213] employed a statistical approach to identify over-represented *D.melanogaster* core promoter motifs using datasets in promoter databases. Gershenson and Ioshikhes [234] employed a statistical approach as well and were able to model potential synergetic combinations in human core promoter motifs. Such studies have advanced understanding of the eukaryotic core promoter architecture by showing that, different transcription initiation modes are fundamentally related to various implementations of basal transcription machinery interactions with promoter DNA. These

interactions are governed by the presence and mutual positioning of core promoter motifs.

## 2.3 A summary of chapter objectives

The overall aim of this chapter was to analyse the core promoters of the newly sequenced *Glossina morsitans* genome. The objectives are outlined below:

- 1) Develop a method to computationally identify TSSs in the recently sequenced *Glossina morsitans* genome.
- 2) Extract core promoters using the experimentally verified TSS locations.
- 3) Analyse properties of core promoters in terms of their TSS distributions as well as core promoter motif composition.



## 2.4 Methodology

### 2.4.1 Acquisition and mapping of *G.morsitans* TSS-seq data

#### 2.4.1.1 Acquisition of *G.morsitans* TSS-seq reads

Briefly, *G.morsitans* pupal and larval samples were prepared at the Yale school of Public health and sent to the sequencing facility at the Genomic Sciences Center in Riken on dry ice. TSS-seq libraries were produced using the oligo-capping method. One-pass sequences were determined using the ABI3730 sequencers. FASTQ files for these larval and pupal TSS-seq reads were downloaded from the DNA Data Bank of Japan experiments SRX004541 and SRX004542 respectively [235]. The files were combined to constitute one file containing approximately 17 million TSS-seq reads.

#### 2.4.1.2 Quality control

The FASTX v 0.13 toolkit [236] was used for read preprocessing by first clipping of adapter sequences using the *fastx\_clipper*. Trimming was performed using the *fastx\_trimmer* with quality filtering by flagging `-t` for minimum quality threshold and `-l` for minimum length of read to be retained after trimming. The `-t` flag was set at 22. The rationale is that, given the maximum quality score is 44, any base with a quality score of 22 and above would give a 50% chance of correct base calling. Several rounds of trimming with variations in `-l` were performed to facilitate optimization of the mapping process. After trimming, the reads were processed with the *fastq\_quality\_filter* with `-q` and `-p` flags set at 31 and 50 respectively. Setting `-q` at 31 ensured that the minimum quality score kept was at least 31 which in effect increased the quality stringency such that only bases with at least 70% probability of correct base calling were retained. In addition a `-p50` threshold ensured that for each read, a minimum of 50% of the read length would have the minimum quality score.

Several rounds of mapping with a range of read lengths (`l=19` to `l=34`) show that

stringent trimming and by extension quality filtering give an improved mapping profile as shown in Figure 2.2. Still, this is a trade-off because an extremely stringent quality filtering procedure eliminates most of the reads (some which may be of good quality).

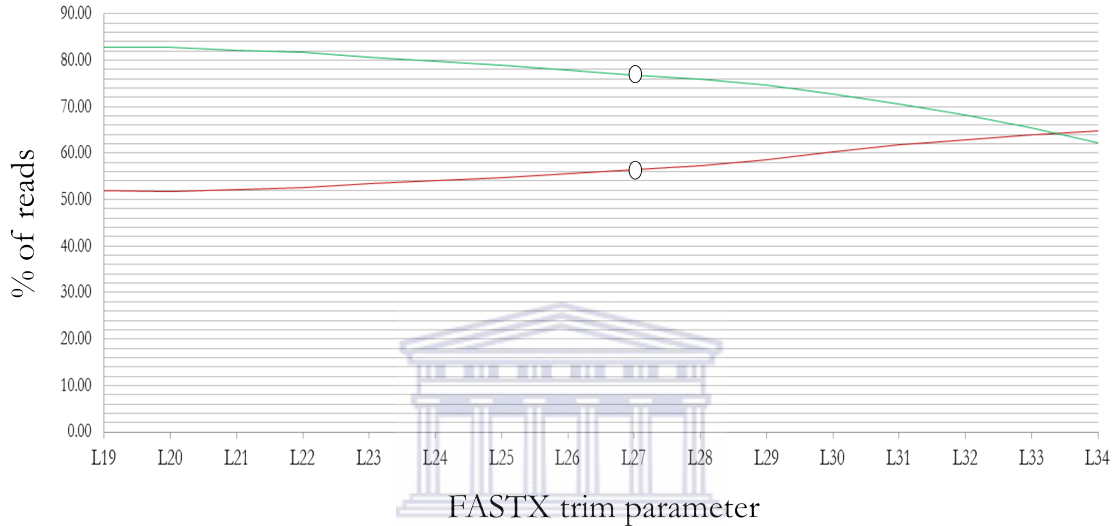


Figure 2.2: Trends of the quality filtering and mapping procedures. The green line represents percent of original reads remaining after quality trimming at different length (L19-L34) variations whilst the red line shows the trend of reads reporting a unique alignment. Small black circles indicate the point L27 that was used as the cut-off point for this analysis.

The trends exhibited by the graph depict a progressive decrease in the percent of original reads reporting a unique alignment after quality trimming such that, the longer the read length higher the % of reads reporting a unique alignment onto the genome (red line). The converse is true for the percent of original reads remaining after quality trimming, the more the trimming the lower the percent of original reads remaining and vice versa. Minimum length for trimming was set at  $l=27$  because at this point at least 70% of the reads from the original dataset of 10 million reads that passed the quality filtering step were retained. In addition, at  $l=27$  at least 60% of the approximately 10 million reads reported a unique alignment onto the genome.

See appendix two for a display of quality charts depicting the change in read quality before and after quality control.

### 2.4.1.3 Mapping

For the reads that passed the quality filtering step, mapping was done using NOVOALIGN [237]. NOVOALIGN unlike other short read mappers, uses the alignment score (flagged using `-t` option) to limit mismatches indirectly. The alignment penalty for mismatches is dependent on the base quality. NOVOALIGN is able to find all the same unambiguous mapping locations nonetheless because it uses base qualities and can align with gaps. Uniquely mapped reads were converted to browser viewable format using BEDTOOLS [238] and uploaded onto GBROWSE [239].

The stringent quality filtering criterion employed in this study reduced the number of reads progressively from the data acquisition through to mapping. Only reads with very high quality were mapped onto the genome. Approximately 41% ( 7 million) of the original (17 million) mapped onto the genome. Out of the 41 %, 60% were reported as unique alignments and were used for subsequent promoter analysis. A summary of these read mapping statistics are shown in Table 2.1 below.

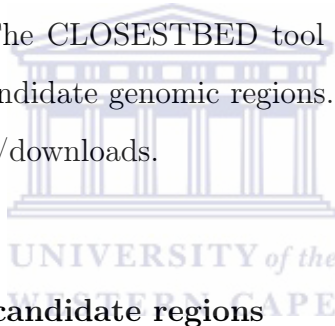
Table 2.1: Summary of read mapping statistics

Parameter	Count
Total number of reads from DDBJ	17,218,719
Reads that passed the quality filtering step	10,543,105
Reads that reported at least one alignment	7,048,660
Reads that reported unique alignment	6,622,424

## 2.4.2 TSS and promoter identification algorithm

### 2.4.2.1 Clustering of TSS-seq reads

The BEDTOOLS suite [238] together with custom scripts were used for the clustering process. Firstly, the alignment of TSS-seq reads was converted from binary alignment (bam) format to browser extendible (bed) format after which the MERGEBED tool was used to merge overlapping reads into a single cluster with `-d` set to zero so that all reads with at least one overlapping base were merged into one cluster. The `-s` option was flagged to force strandedness and `-n` to obtain a count of the reads that were contained in the corresponding clusters. For core promoter analysis, clusters that contained at least 100 TSS-seq reads were selected. They are referred to as tag clusters henceforth. The CLOSESTBED tool was used to classify tag cluster's mapping positions on candidate genomic regions. These scripts are available on <http://gmpromdb.sanbi.ac.za/downloads>.



### 2.4.2.2 Identification of candidate regions

A genic region was defined as a region spanning the gene from the 5'UTR through to the 3'UTR. Classification of clusters mapping onto the 5'UTR was done as follows;

- (i) *Bona fide* 5'UTR clusters = tag clusters mapping onto annotated 5'UTR regions.
- (ii) Other 5'UTR clusters = tag clusters mapping on genes for whom no 5'UTR is annotated.

For the 'other 5'UTR' category, summary statistics for the 5'UTR lengths were computed. Tag cluster(s) mapped to a maximum of 300 bases (mean 5' UTR length) from the start of the first coding exon were included (Figure 2.3). In addition, genes with extremely short (1-10 bases) predicted 5'UTRs but with a tag cluster(s) mapped to a maximum 300 bases (mean 5' UTR length) from the start of the 5'UTR were also included. We did not exclude the TSS that mapped in the first coding exons because it has been shown that the locations of TSS fluctuate to some extent in most



genes [72] [73] [240].

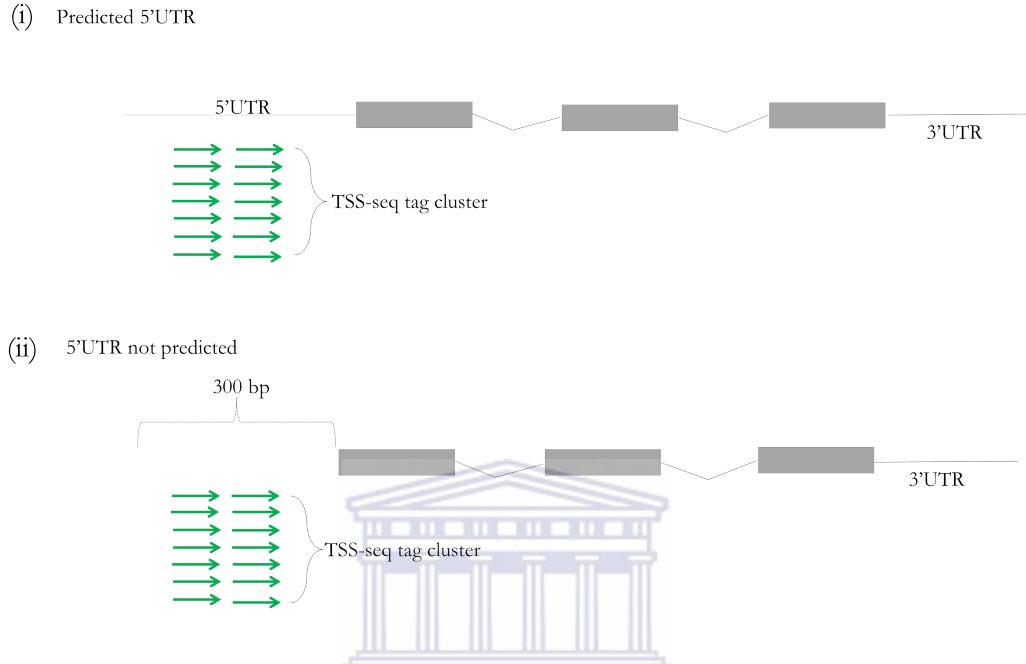


Figure 2.3: Representation of the classification of 5'UTR tag clusters. (i) represents *bona fide* 5'UTRs while (ii) represents genes for which the 5'UTR was lacking. For the latter, tag clusters mapping within 300 bases of the first coding exon were captured for promoter extraction.

Before generating the final tag clusters data set, tag clusters that had been mapped inside the second or later exon of the gene models were excluded from further analysis. It has been suggested that these TSS would be artifacts arising from recapped transcripts [92].

A custom script was used to extract and calculate the frequency of tag start positions for each cluster. A TSS position was defined as the position with the highest frequency of tag counts for each delineated cluster. A summary of the TSS-seq mapping and tag clustering protocols are outlined in Figure 2.4 below.

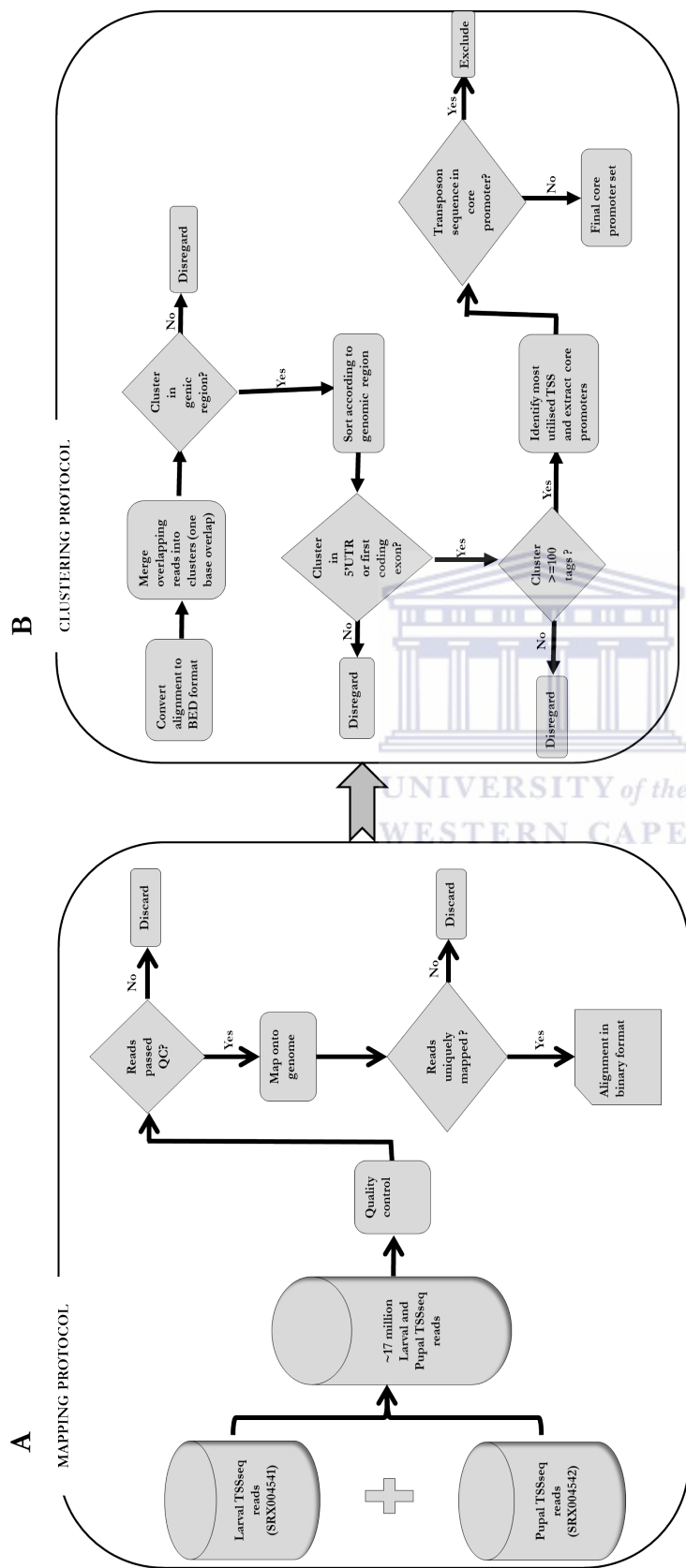


Figure 2.4: Workflows summarizing the mapping and clustering protocols. (A) The mapping protocol illustrates the procedure from acquisition of TSS-seq reads, quality control, mapping and capturing of unique alignments. (B) Clustering protocol illustrates the procedure from clustering of TSS-seq reads, sorting according to genomic region, identification of most utilised TSS and extraction of core promoters.

### 2.4.2.3 Delineation of promoter classes

Core promoters were defined as one hundred bases in the milieu of the TSS (-50/+50 relative to the TSS (+1)). Some broad core promoters exhibit properties of both narrow and broad initiation patterns where they exhibit propensity for one TSS. These are classified as “broad with peak” while those that do not exhibit propensity for one particular TSS are referred to as “broad without peak” [241]. To delineate the shapes (with or without peak) of a tag cluster, we employed the individual peakedness score method as described by Zhao and colleagues [242]. This method evaluates the peakedness of tag clusters by defining the individual peakedness “ $s$ ” of a tag cluster “ $g$ ” by the formula;

$$Sg = \frac{m}{nw} \quad \text{equation(2.1)}$$

Where  $m$  is the tag count at the dominant peak (the mode)  $n$  is the total number of reads in the distribution,  $w$  is the width of the distribution, that is, the genomic window covered by the tag cluster.

For illustration purposes, Figure 2.5 below is an example of a TSS distribution for a cluster in the dataset. The figure can be loosely defined as a histogram of tag counts along the 5' UTR or the first coding exon of the corresponding gene in the genome.

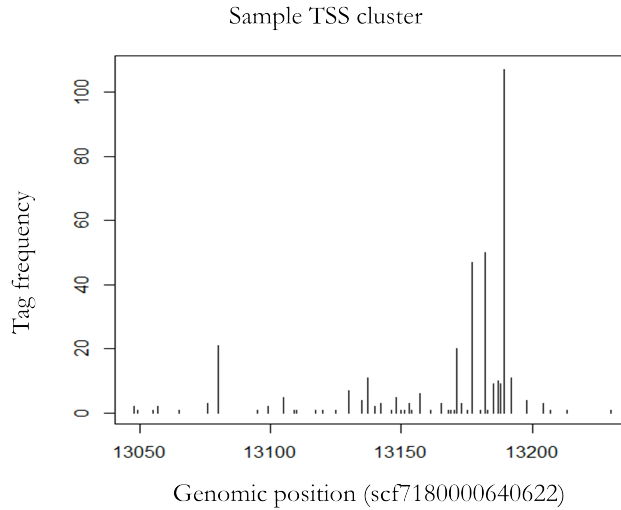


Figure 2.5: Histogram of a sample TSS-seq tag cluster. The x axis represents the genomic TSS positions where TSS-seq reads map (the corresponding scaffold ID is denoted) while the y axis represents the frequency of reads at each TSS.

Substituting the values according to equation 2.1;

The mode (m) (Number of TSS-seq reads) = 107

There are 49 TSS positions, thus the width (w) = 49

The sum (n) of the reads (counts) is:

$$(9+10+1+50+1+9+107+1+2+1+3+1+5+21+1+2+1+1+1+1+1+5+1+3+1+3+1+4+1+6+1+3+1+3+2+11+1+4+11+1+1+7+20+1+3+1+2+47+1)=375$$

The individual peakedness score will be  $107/(375*49) = 0.005823129$

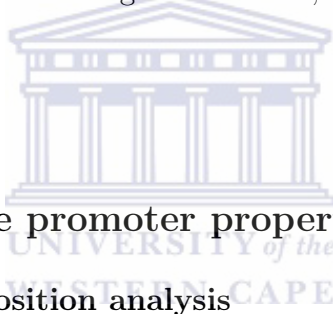
Accordingly, each cluster will have its discrete individual peakedness score. Notably the maximum individual peakedness score will have a value of 1, where a cluster has only one TSS position. These are referred to as peaked clusters. The higher the individual peakedness score, the more defined the TSS genomic location.

The absolute count of TSS positions in a given tag cluster was used to classify broad versus narrow promoters. All clusters with 10 or less TSS positions were classified as narrow whilst the remainder was classified as broad. Broad clusters were further

categorized based on their peakedness. A broad cluster was deemed as one with a dominant peak if the TSS with the highest frequency of tag counts constituted at least 50% of the total tag count. If this condition was not met, the broad cluster was classified as “without a peak”.

#### **2.4.2.4 Core promoter extraction**

Core promoters were defined as 100 nucleotides (-50/+50) surrounding the TSS. An in house script was used to extract the core promoters. For candidate regions with more than one tag cluster, the cluster with more tags was selected for further analysis. To avoid confounding results during motif search, core promoters entangled with transposons were eliminated.



### **2.4.3 Analysis of core promoter properties**

#### **2.4.3.1 Nucleotide composition analysis**

To compare the nucleotide composition between mammalian and insect genomes, human and *Drosophila* promoter datasets utilized in the study by FitzGerald and colleagues [70] were obtained. The mouse promoter dataset was obtained from the Eukaryotic promoter database [243]. Regions spanning the TSS from -200 to +100 positions were extracted to assess percent nucleotide composition at each position using the FASTX toolkit [236].

#### **2.4.3.2 Annotation of core promoter motifs**

The canonical RNAPol II mediated core promoter motifs consist of; the TFIIB recognition motifs BREu and BREd, the TATA-box, INR, DPE and MTE. In the absence of experimentally verified core promoter motifs for *G.morsitans* hitherto, core promoter associated PWMs were downloaded from the JASPAR database [244]. MATRIX-SCAN program [245] of the RSAT suite [246] was used to scan core promoter se-

quences. MATRIX-SCAN scans sequences with one or several PWMs to identify instances of the corresponding motifs. A score is assigned to each position on the candidate sequence based on its similarity with the motif described in the PWM. Only predictions that reach some predefined threshold are retained as predicted binding sites. MATRIX-SCAN estimates a p value for each site. The p value evaluates the probability to obtain a given score by chance thereby enabling removal of potential false positive predictions. The p value threshold was set at  $10e^{02}$  to distinguish true positives from false positives and the program was set to estimate a background model from the input sequence. Only those core promoter motifs occurring at their corresponding biologically functional genomic windows were captured. Because core promoter motif placement often exhibits elasticity [78], the canonical start positions were allowed to vary by +/-5 bp within their corresponding genomic windows for which they are biologically functional. These windows are tabulated below.

Table 2.2: Core promoter genomic windows used for the analysis

Core promoter motif	Genomic window for motif start
BREu	-37 to -27
TATA-box	-31 to -21
BREd	-23 to -13
INR	-5 to +5
MTE	+18 to +28
DPE	+28 to +38

Because TFBSs are typically short (5-15 nucleotides) and tolerate generally high levels of sequence degeneracy, majority of common motif models may not accurately discriminate *bona fide* motifs from remaining sequence. Indeed, every motif finding algorithm is usually exposed to spurious matches that may appear as significant as the ones in questio. This aspect has been described as the motif “twilight zone” [247]. By knowing the biologically functional genomic windows of these motifs *a priori*,

motifs that did not occur in biologically functional genomic windows were excluded, partly correcting for spurious matches. Further, random promoter sets for each of the promoter classification that is; narrow, broad with peak and broad without peak were generated by shuffling the core promoter sequences. A similar technique was employed by Frith *et al.*, [248] and Jin *et al.*, [66]. Though non-functional, the random promoter sets have the same nucleotide frequencies as the true core promoter sequences. Accordingly, they are preferable to coding or inter-genic sequences which exhibit nucleotide bias. The core promoter motif annotation procedure was repeated for the random promoter datasets. A paired binomial test was performed to determine whether there were significant differences in the numbers of core promoters harboring motifs at biologically functional genomic windows between the *bona fide* and randomly generated core promoter datasets.

The main weakness of this approach is that it is simulation based. In essence simulations greatly facilitate assessment of significance of any motif finding result, but they do not show how the false positive rate changes as the motif finding parameters are altered. Additionally, given that this is a newly sequenced genome with the manual curation of the computationally predicted gene models in progress, estimation of motif occurrence depend on the quality of the gene models as well as the cut-off parameters. Different cut off values would produce varying results. However, we are of the conviction that significant p values from the binomial tests would be a good indicator of the reliability of our predictions.

#### **2.4.3.3 Generation of *Glossina* specific core promoter PWMs**

The MEME program [249] was used to build *G.morsitans* specific core promoter motif matrices. For each core promoter motif identified by MATRIX-SCAN, regions spanning the biologically relevant windows were extracted and used as input for MEME. Briefly, MEME takes a group of DNA and applies statistical modeling techniques to sequences and outputs as many motifs as requested. The program automatically

chooses the number of occurrences and description for each motif as well as the best width. MEME was run with the parameters `-mod oops -nmotifs 1 -DNA -revcomp`. The output of MEME is motifs that are represented as PWMs. These PWMs describe the probability of each possible letter at each position in the pattern. Each MEME motif is devoid of gaps. Patterns with gaps of variable lengths are split by MEME into two or more independent motifs.

The output from the MEME program with the lowest p value and highest number of hits was selected after which the motifs were used to search for known experimentally verified RNA pol II TFBSs using TOMTOM [417]. TOMTOM matches an input DNA motif with the motifs of a database of known motifs together with their corresponding reverse complements. Matching motifs are compiled into a list and reported. This list ranked by q value which denotes the minimal false discovery rate at which the observed similarity would be considered significant. For a given pair of motifs, the program takes all offsets into consideration while requiring a minimum number of overlapping positions. Each overlapping position for a given offset is scored using similarity function. Scores of columns that overlap for each offset are summed and converted to a p value. The minimal p value amongst all possible offsets is reported. The p values are then changed to q values for each query motif [417]. *Glossina* specific core promoter matrices together with their corresponding logos are attached in appendix three.

#### **2.4.3.4 Analysis of core promoter motif co-occurrence**

Motif co-occurrence suggests combinatorial regulation of transcription via physical interactions between corresponding TFs. To elucidate patterns of motif co-occurrence in various core promoter classes, two-way and three-way motif co-occurrences were evaluated for each core promoter sequence. Further, core promoter sequences with a TATA-box in a biologically functional window (TATA-containing) were separated from core promoters without a TATA-box in a biologically functional window (TATA-



less). The frequency of motifs in TATA-containing versus TATA-less categories was also evaluated. A summary of the core promoter analysis protocols is shown in figure 2.6 below.

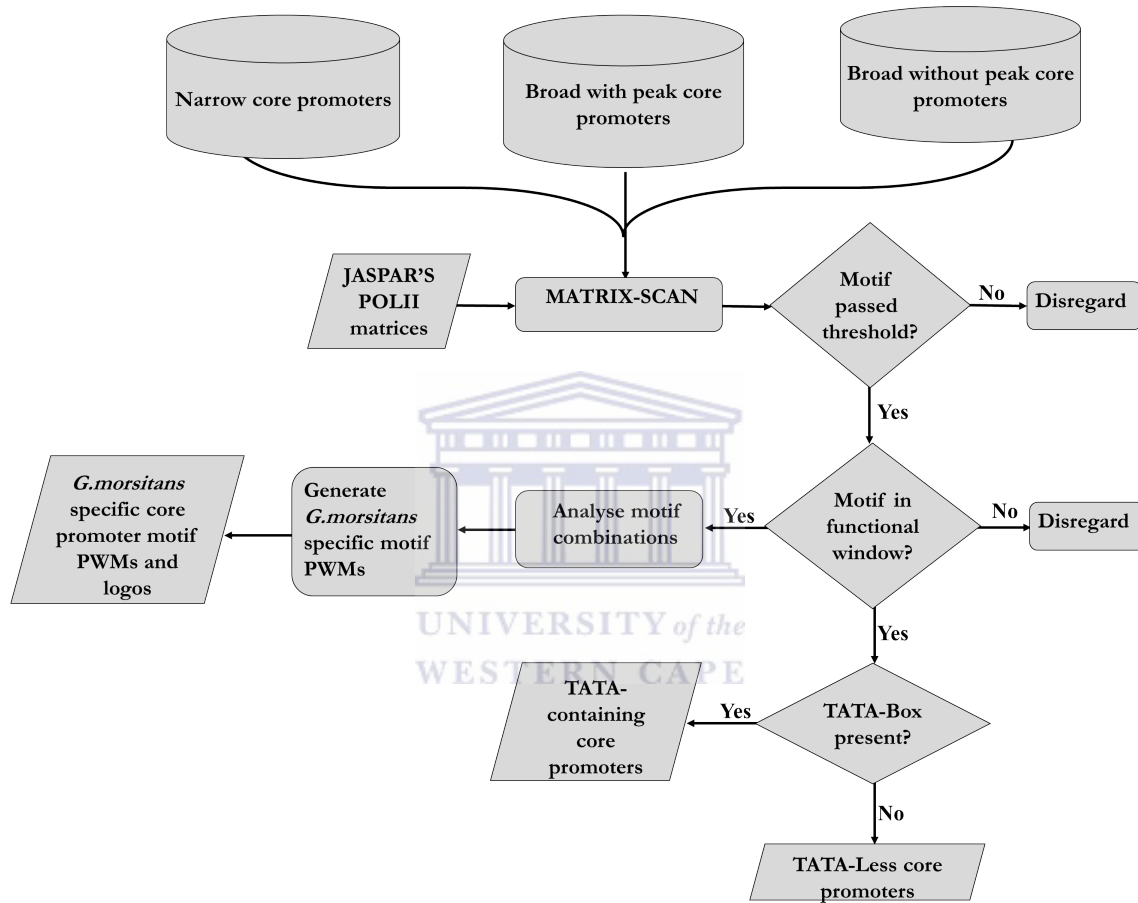


Figure 2.6: Summary of core promoter motif analysis methodology.

#### 2.4.4 Analysis of functional aspects of different promoter classes

The ability of TSS-seq to facilitate digital expression profiling was exploited. It was hypothesized that gene products characteristic of development would be over-represented since larvae and pupae samples were used to create TSS-seq libraries. To evaluate for overrepresentation of certain biological/molecular processes in the three promoter classes, Gene Ontology (GO) [251] annotations were obtained for every gene in each promoter classification. For each promoter class, GO annotations were

summed and the summary statistics computed. Annotations falling within the 75th percentile of their corresponding summary statistics were deemed as over-represented. A summary of this methodology is shown in Figure 2.7 below.

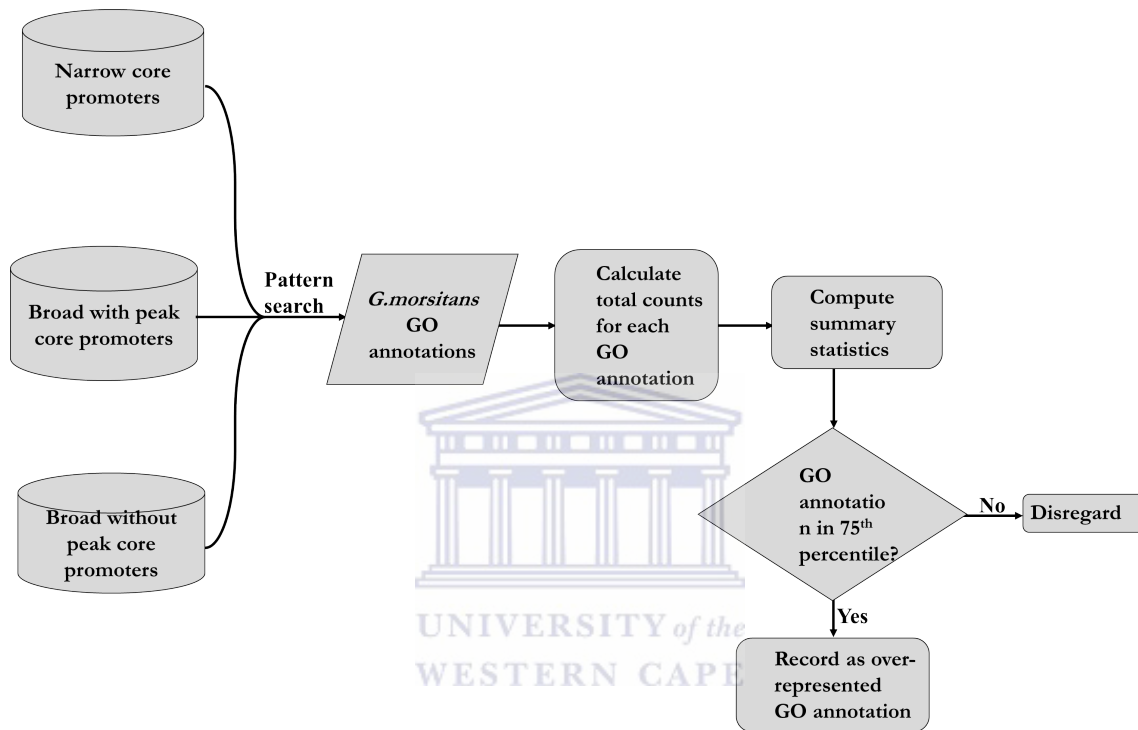


Figure 2.7: Summary of core promoter GO annotations analysis methodology.

## 2.5 Results

### 2.5.1 Genome mapping statistics

The current assembly of the *G.morsitans* genome consists of 13,807 scaffolds 3,058 of which contain at least one gene. TSS-seq reads map onto 2,896 of the 3,058 gene-containing scaffolds and 2,736 out of 10,749 gene-less scaffolds. There is an almost equal number of gene-containing versus gene-less scaffolds with tag clusters, (2,896 and 2,736 respectively). However, majority of the tag clusters mapped onto gene-containing scaffolds (Table 2.3 actual numbers are highlighted with an asterisk). Some gene-less scaffolds may be intergenic regions as the genome is yet to be reconstructed into its karyotype.

Table 2.3: Summary of genome mapping statistics

Parameter	Count
Gene-containing scaffolds	3058
Gene-containing scaffolds with TSS-seq reads	2,896 (6,513,739*)
Gene-less scaffolds	10,749
Gene-less scaffolds with TSS-seq reads	2,736 (108,685*)

\*Absolute numbers of mapped reads.

### 2.5.2 Clustering statistics

A total of 3134 tag clusters were obtained. Their distribution is described in the table below;

Table 2.4: Summary of tag clustering statistics

Parameter	Number of clusters
Number of clusters with at least 100 TSS-seq reads	3134
Total number of tag clusters in genic region	2033
Total number of tag clusters outside genic regions	1014
Total number of tag clusters in CDS1* and 5UTR	1424
Total number of tag clusters in other genic regions	609
Tag clusters in non-gene containing scaffolds	87

\*CDS1=First coding exon

Approximately 65% (2033/3134) tag clusters mapped onto the genic regions while 35% (1101/3134) tag clusters mapped onto the intergenic regions, and may represent previously unannotated transcripts or non-coding RNA TSS. 70% (1424/2033) tag clusters that mapped onto the genic region were located on either the first coding exons or their proximal 5'UTR regions. An additional 87 tag clusters mapped onto the gene-less scaffolds that may also represent intergenic regions.

The phrase ‘intergenic cluster’ is used with caution because of the fragmented nature of the genome coupled with irregular gene distributions. For instance, approximately two-thirds of the gene-containing scaffolds harbor only one gene. Furthermore, the median gene length was found to be 4488 base pairs. Tag clusters located at most 4488 base pairs from either end of the scaffold were classified as “periphery tag clusters”. These tag clusters could not be associated with any gene.

Our calculations established that the mean length of predicted 5' UTR for the current genome assembly is 300 bases. For genes without a predicted 5'UTR tag, clusters with a least ten reads and mapping not more than 300 bases from the start of the first coding exon of these genes were identified. This resulted in identification of putative TSS for approximately 200 genes, thus aiding in the annotation of 5'UTRs for these

genes.

### 2.5.3 Delineation of promoter classes

The peakedness of a TSS tag cluster decreases with size (see Figure 2.8). Clusters with a single TSS had a  $Sg$  value of 1 whilst the cluster with the highest number of TSS (220) had 0.0023 as its  $Sg$  value. Essentially higher  $Sg$  values represent peaked distributions a while low  $Sg$  values represent non-peaked distributions as shown in Figure 2.8.

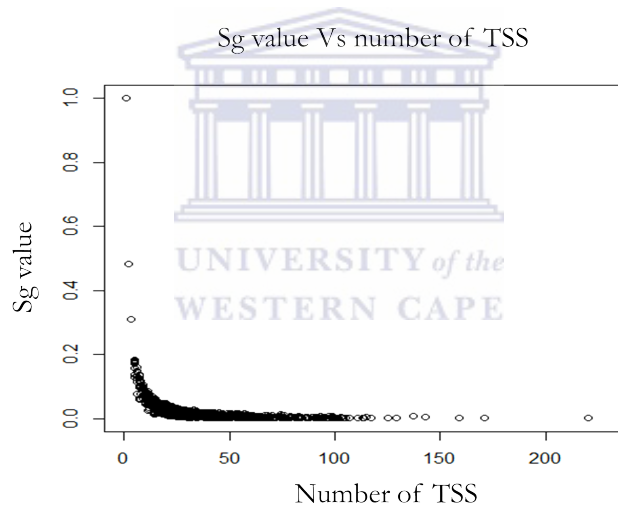


Figure 2.8: Relationship between  $Sg$  value and number of TSS positions. The graph depicts a negative correlation between the number of TSS positions and the  $Sg$  value.

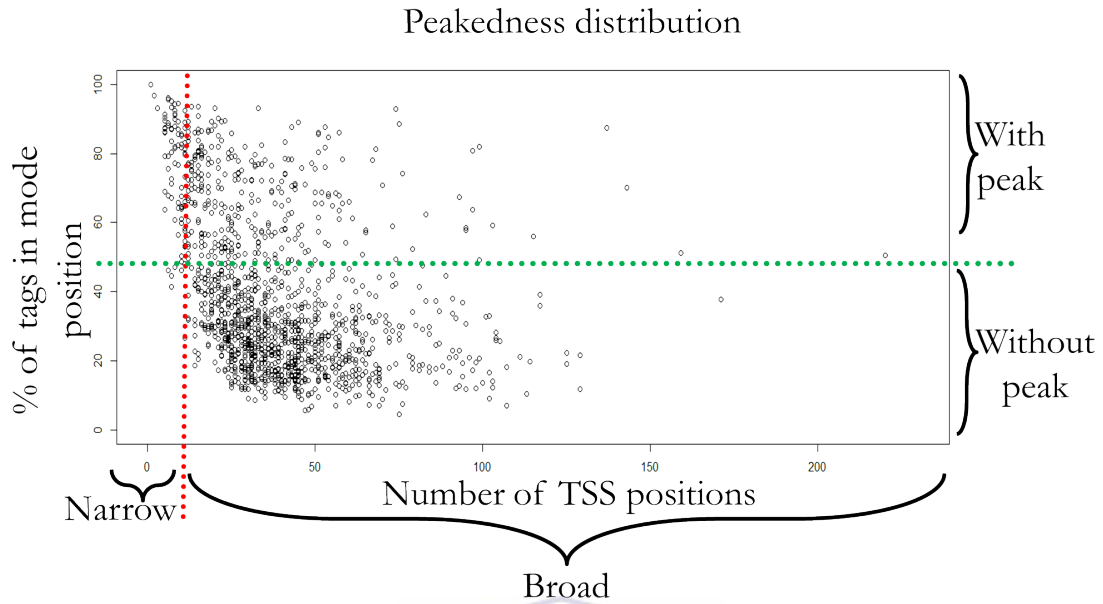


Figure 2.9: Relationship between number of TSS positions (x axis) and percent of tag count at the mode position value (y axis). Red line shows the borderline between narrow and broad tag clusters whilst the green line denotes the borderline between broad with peak and broad without peak tag clusters.

The region with a high point density illustrates that majority of the tag clusters are of the broad without peak category.

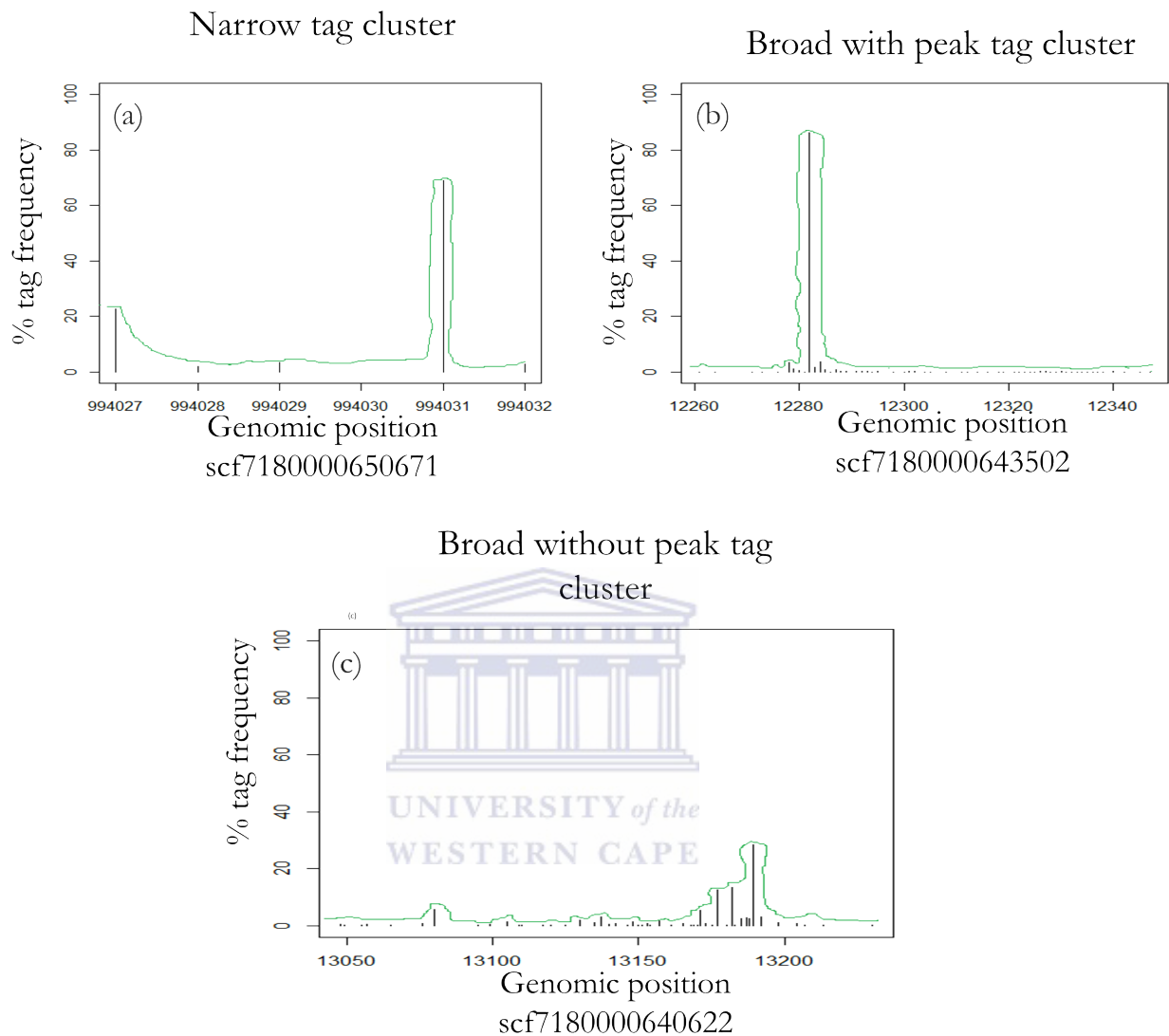


Figure 2.10: Graphical impressions of representative tag clusters for the various promoter classes. The x axis represents the genomic position. The corresponding scaffold ID is denoted. The y axis represents the percent of the total tag count at each genomic position. Figure (a) is a representative of the narrow class whose TSS positions span five nucleotides with a single dominant peak. Figures b and c denote the broad promoter classes whose TSS spans several to hundreds of nucleotides. This class can either be broad with a dominant peak (b) where the dominant peak constitutes approximately 80% of the total tag count or broad with multiple but no dominant peaks(c) where there is no single dominant peak.

The cutoff of at least 50% of the total reads in the mode position for a tag cluster to be classified as peaked resulted in more than 70% of the tag clusters without peaks. The results are summarised below;

Table 2.5: Summary of tag cluster types

Tag cluster type	Absolute number	% of total
Narrow	69	5
Broad with peak	314	23
Broad without peak	1010	72

#### 2.5.4 Core promoter extraction and classification

Core promoters were defined as 100 nucleotides (-50/+50) surrounding the TSS. By setting the threshold at 100 tags per cluster, 1424 tag clusters were located on the first coding exons and their proximal putative 5'UTR regions. Where a gene had more than one candidate tag cluster, the cluster with more tags was selected for further analysis. Their corresponding core promoters were extracted. Approximately 31 core promoters were entangled with transposons and were therefore excluded from further analysis. The core promoter set used for motif assignment was therefore reduced to 1393.

#### 2.5.5 Comparison of core promoter nucleotide distribution

Alignment of -200/+100 regions surrounding the TSS showed that insect promoters exhibit propensity for the AT dinucleotide whilst mammalian promoters exhibit propensity for the CG dinucleotides as shown in the charts below.



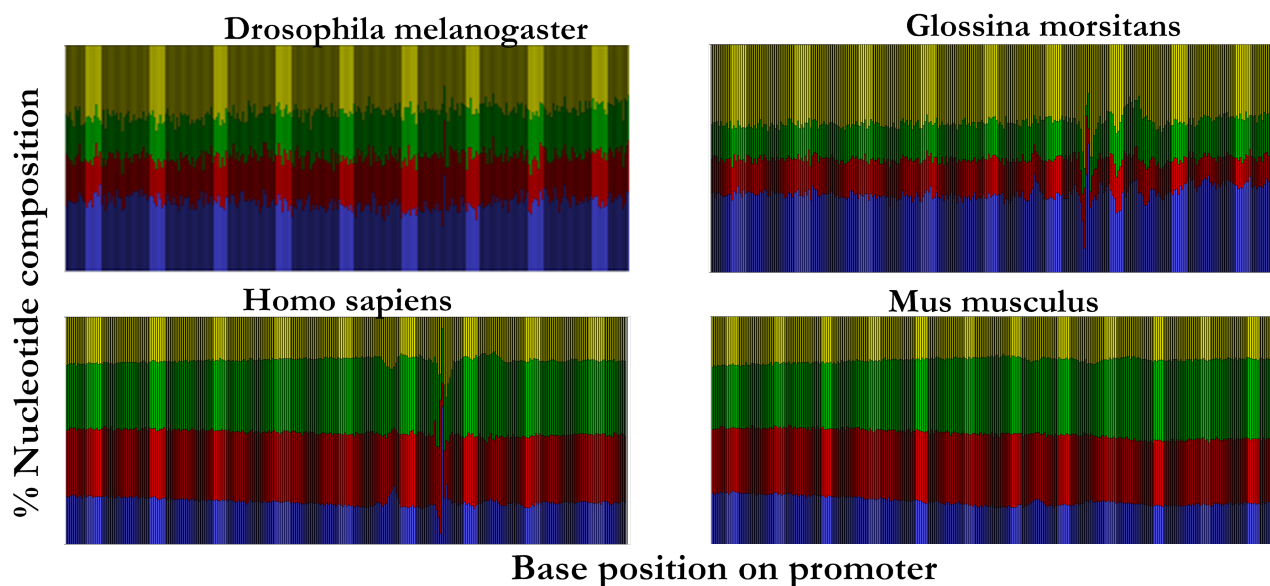


Figure 2.11: Combined promoter nucleotide composition graphs for *D.melanogaster*, *G.morsitans*, *H.sapiens* and *M.musculus*. The regions represented encompass the -200 to +100 regions flanking the TSS. The y axis represents % base composition at each nucleotide position. The x axis represents 300 nucleotides in the TSS milieu. Red=C, Blue=A, Green=G, Yellow=T.

The charts above show that there is clear distinction in nucleotide composition between mammalian and insect promoters. While the insect promoters exhibit propensity for the AT nucleotides, mammalian promoters exhibit propensity for the CG nucleotides.

The CA dinucleotide often associated with the TSS [252] but it has recently been shown that the INR pattern varies substantially between studies, ranging from a TCA (G/T) TC(C/T) to a single dinucleotide (pyrimidine (C/T)–purine (A/G)) [53]. Notably, majority of promoters only have one or a few of these patterns, and some patterns are typically found in certain species.

The -5/+5 region surrounding the TSS was extracted for closer scrutiny of the INR in *G.morsitans* core promoters (Figure 2.12).

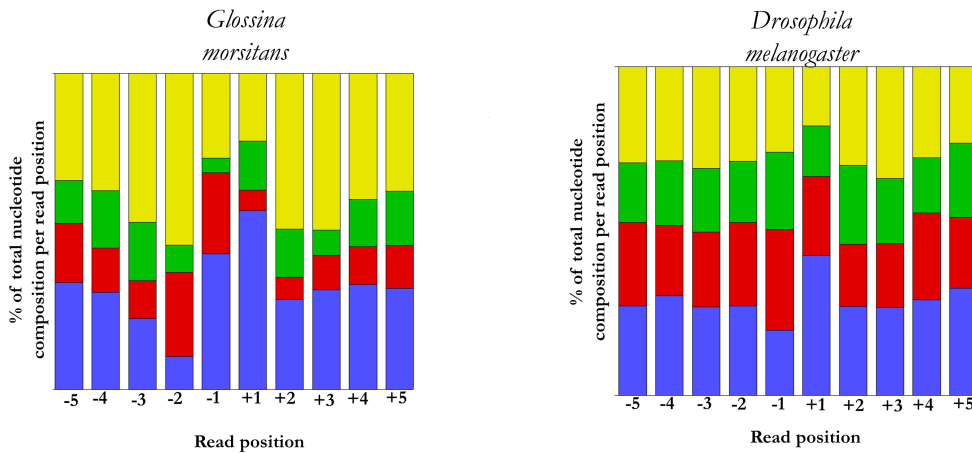


Figure 2.12: *G.morsitans* core promoter’s nucleotide frequency surrounding the TSS. The graph represents the region -5+5 nucleotides around the TSS. The y axis represents % base composition at each nucleotide position. The x axis represents 10 nucleotides surrounding the TSS. Red= C, Blue=A, Green=G, Yellow=T

The nucleotide frequency distribution shows that in *G.morsitans* core promoters, the -1/+1 positions show propensity for the AA dinucleotide while the corresponding positions in *D.melanogaster* core promoters show propensity for the CA dinucleotide.

## 2.5.6 Annotation of core promoter motifs

### 2.5.6.1 Real vs random motifs

In all cases the number of motifs identified within the real promoter datasets exceeded the number found in the randomized sequences, indicating a positive signal for the core promoter motifs. Narrow core promoters had a p value of 0.00164, while broad with peak core promoters had a p value of 0.00135, the p value for broad without peak promoters was 0.00185. Table 2.6 below shows the absolute numbers of core promoter motifs between the true and random datasets.

Table 2.6: Comparison of core promoter motif instances between true and random datasets

Core promoter category	BREu	BREd	TATA	INR	MTE	DPE
*(a) Narrow (true)	15	13	27	28	16	20
Narrow(random)	9	5	10	10	13	5
*(b) Broad with peak (true)	38	61	88	87	74	80
Broad with peak (random)	32	33	45	42	39	47
*(c) Broad without peak (true)	129	180	163	181	190	166
Broad without peak (random)	103	106	131	133	102	102

\*(a) p value 0.00164, \*(b) p value 0.00135, \*(c) p value 0.00185.

As mentioned before, every motif finding algorithm is ordinarily exposed to spurious matches that may appear as significant as the ones in question. The numbers of core promoter motifs in the random dataset confirm this (Table 2.6). However, we know *a priori* that the real dataset consists of functional promoters. Obtaining p values that were below the predefined threshold that is 0.05, (95% confidence interval) indicates that the occurrence of core promoter motifs in the biologically functional window for *G.morsitans* core promoters is enriched in the true dataset compared to the random dataset.

### 2.5.6.2 Core promoter motifs in narrow versus broad classes

Canonical core promoter motifs were found in approximately 74% of *G.morsitans* core promoters. Figure 2.13 shows a clear separation in core motif frequency between narrow and broad promoters. While the BREu, TATA, INR and DPE were more prevalent in narrow promoters, broad promoters exhibited propensity for the MTE and BREd.

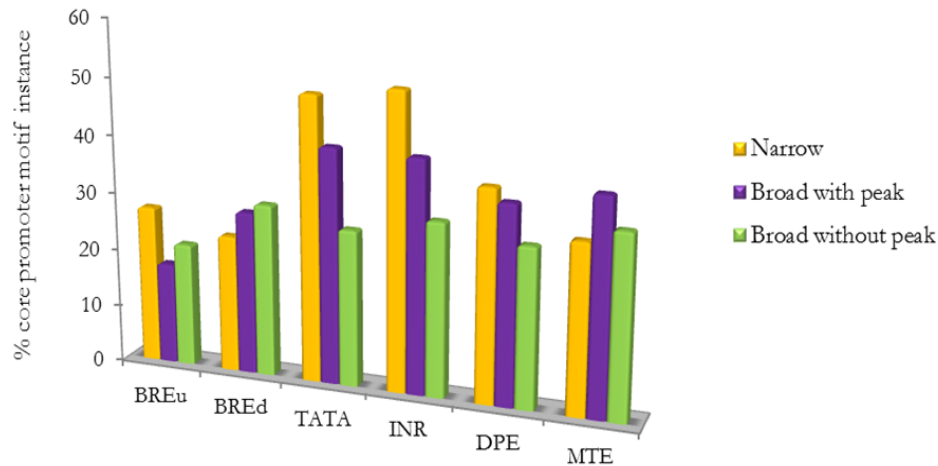


Figure 2.13: Graphical summary of core promoter instances in the various promoter classes.

The greatest variation in core promoter motif frequency between narrow and broad promoters was observed for the TATA-box and INR motifs; approximately 50% of the narrow core promoters harbored these motifs. This suggests that the INR may be of equal importance to transcription for narrow promoters as the TATA-box in *G.morsitans*. Narrow core promoters with focused initiation sites are associated with motifs such as the TATA-box and INR [214].

This study indicates that 23% of *G.morsitans* core promoters used for this analysis harbor a TATA-box. There is no significant difference in the frequency of the BREu motif; however, the remaining core promoter motifs have a higher frequency in TATA-less promoters, notably the MTE and INR motifs, see Figure 2.14 below.

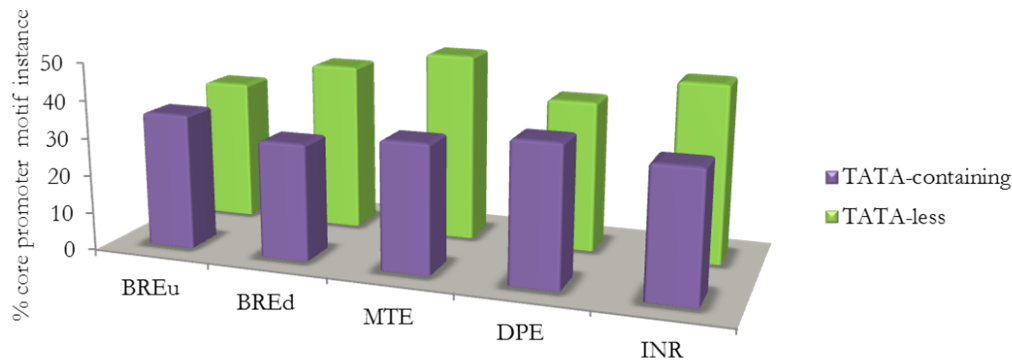
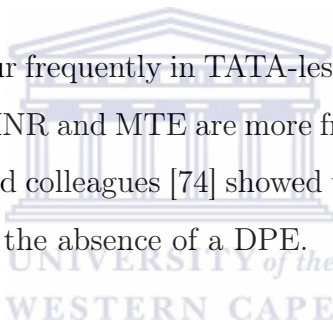


Figure 2.14: Graphical summary of core promoter instances in the TATA containing and TATA-less core promoters.

The DPE was reported to occur frequently in TATA-less promoters in *Drosophila* [75], but this results show that the INR and MTE are more frequent in *G.morsitans* TATA-less promoters. Indeed Lim and colleagues [74] showed that the MTE can compensate for the loss of a TATA-box in the absence of a DPE.



### 2.5.6.3 Motif combinations in narrow vs broad core promoters

Essentially, positive associations between motifs suggest possibility of physical interactions between the TFs that bind the co-occurring motifs. Negative correlations on the other hand imply that the TFs that bind them have divergent functions. The percent two-way and three way motif co-occurrence was computed for each core promoter class producing fifteen possible two-way interactions and twenty possible three-way interactions. For each core promoter category, the pair/triplet with the highest percent co-occurrence was identified.

Narrow core promoters exhibit propensity for the TATA-INR pair (Table 2.7 green shaded area). Broad with peak core promoters exhibit propensity for the INR-MTE and MTE-DPE (Table 2.7 cyan and magenta shaded areas respectively) and MTE-DPE pair. The broad without peak core promoters show a propensity for the MTE-

DPE pair (Table 2.7 yellow shaded area).

Table 2.7: Two way motif co-occurrences

% co-occurrence in core promoter category			
Motif combination	Narrow	Broad with peak	Broad without peak
BREu-TATA	14	8	13
BREu-BREd	12	13	15
BREu-INR	19	13	10
BREu-MTE	7	12	10
BREu-DPE	17	11	7
TATA-BREd	21	13	11
TATA-INR	28	14	10
TATA-MTE	23	13	9
TATA-DPE	27	13	13
BREd-INR	11	13	12
BREd-MTE	12	13	11
BREd-DPE	10	9	11
INR-MTE	22	20	10
INR-DPE	14	17	12
MTE-DPE	20	20	22

Three-way motif co-occurrence indicated distinct combinations for each of the core promoter classes. The TATA-MTE-DPE triplet is the preferred combination for narrow core promoters (Table 2.8 green shaded area) whilst broad with peak core promoters show preference for TATA-INR-MTE (Table 2.8 magenta shaded area), and TATA-INR-DPE triplet (Table 2.8 yellow shaded area). Broad without peak core promoters exhibit propensity for the INR-MTE-DPE triplet (Table 2.8 cyan shaded area).

Table 2.8: Three way motif co-occurrences

% co-occurrence in core promoter category			
Motif combination	Narrow	Broad with peak	Broad without peak
BREu-TATA-BREd	26	23	27
BREu-TATA-INR	42	24	22
BREu-TATA-MTE	35	24	25
BREu-TATA-DPE	31	22	21
BREu-BREd-INR	40	24	22
BREu-BREd-MTE	22	24	20
BREu-BREd-DPE	24	17	21
BREu-INR-MTE	35	27	23
BREu-INR-DPE	35	29	19
BREu-MTE-DPE	28	31	22
TATA-BREd-INR	40	30	22
TATA-BREd-MTE	35	27	23
TATA-BREd-DPE	33	30	22
TATA-INR-MTE	26	36	25
TATA-INR-DPE	44	36	23
TATA-MTE-DPE	49	35	28
BREd-INR-MTE	20	31	23
BREd-INR-DPE	27	30	21
BREd-MTE-DPE	32	30	24
INR-MTE-DPE	40	32	29

### 2.5.7 Functional classification of different promoter classes

Ontology terms were assigned to 307 out of 1393 that represented different promoter classes. Selection of ontology terms in the 75th percentile of each promoter category showed that ontologies associated with developmental processes such as structural

constituents of cuticle are present in all core promoter classes (Figures 2.15-2.17). Cuticular constituents are involved in chitin metabolism and molting and are crucial to insect growth and morphogenesis [253]. Other frequently occurring ontologies include structural constituent of ribosome, small GTPase mediated signal transduction and heat shock protein binding. Other ontology terms largely constituted signaling pathways that may facilitate *G.morsitans* developmental processes (Figures 2.15-2.17).

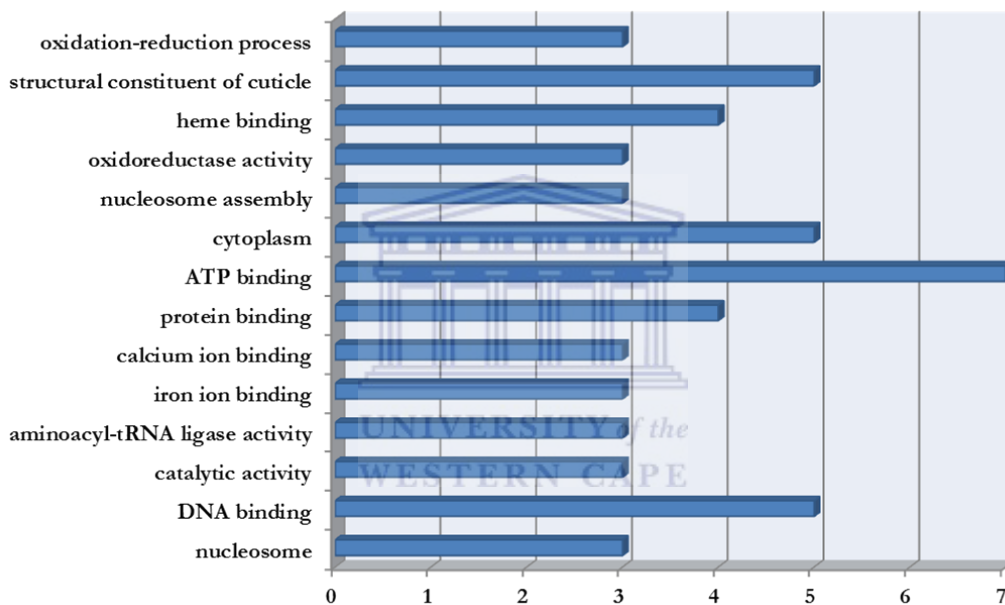


Figure 2.15: Ontology terms occurring in the 75th percentile of narrow core promoters. In the narrow category, the ontology terms structural constituent of cuticle, ATP binding and DNA binding recorded highest frequency.



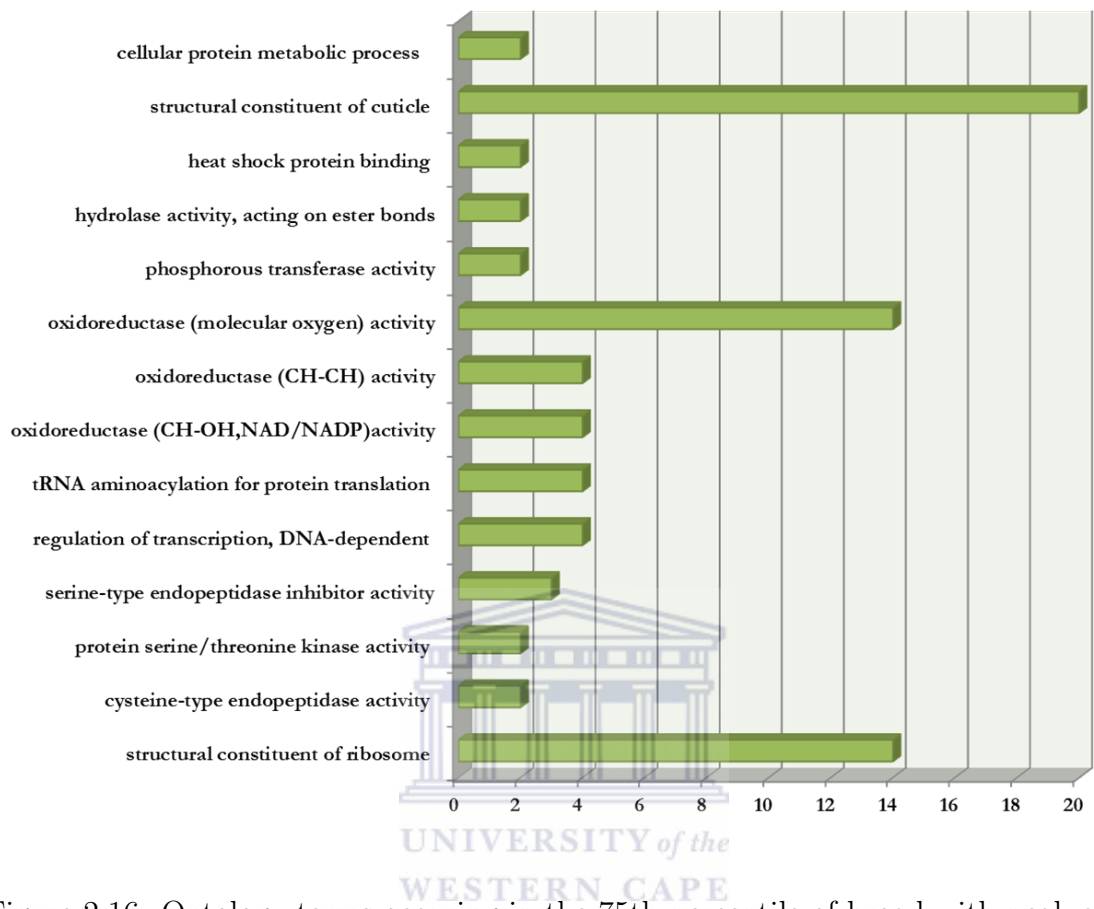


Figure 2.16: Ontology terms occurring in the 75th percentile of broad with peak core promoters. In the broad with peak category, the ontology terms structural constituent of cuticle, oxidoreductase with molecular oxygen activity and structural constituent of ribosome recorded highest frequency.

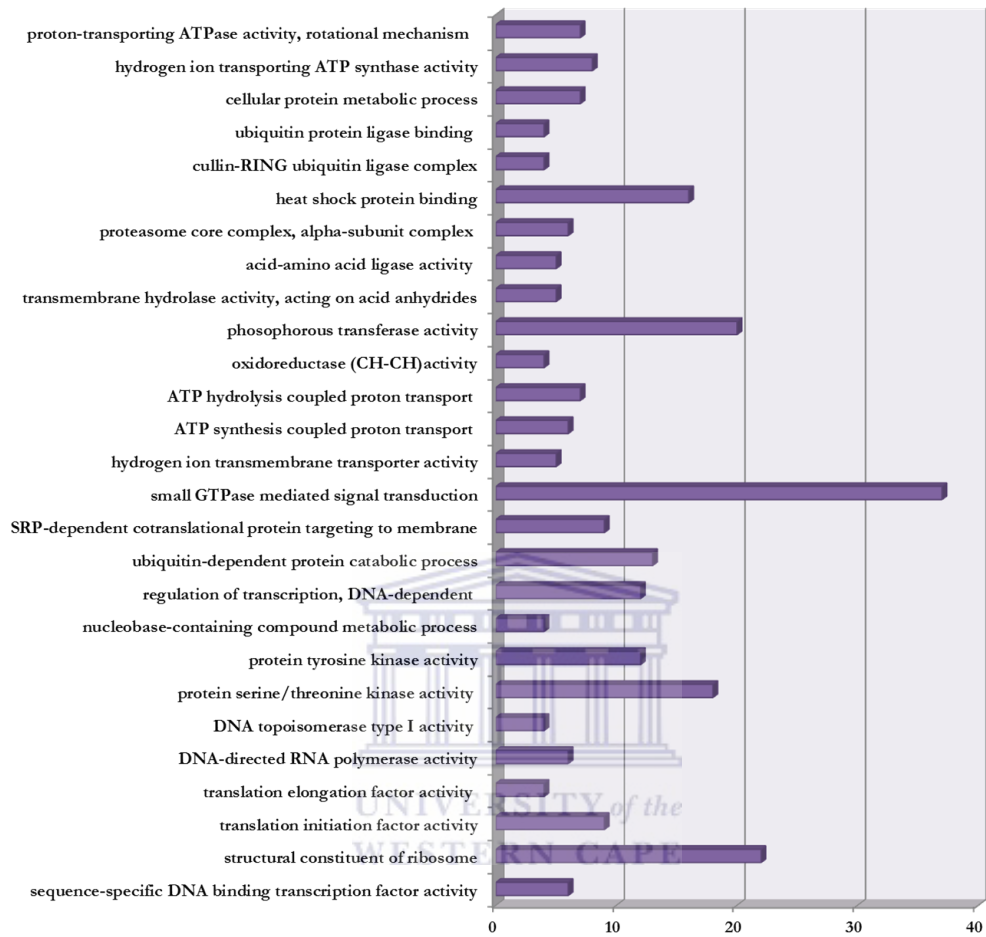


Figure 2.17: Ontology terms occurring in the 75th percentile of broad with peak core promoters. In the broad without peak category, the ontology terms with high frequency include: small GTPase mediated signal transduction, structural constituent of ribosome, protein serine/threonine kinase activity, structural constituent of ribosome, phosphorus transferase activity and heat shock binding activity binding.

## 2.6 Discussion

**Similar to other metazoans, the *G.morsitans* transcriptional control program is characterized by both narrow and broad promoters.**

It was presumed that the combinatorial interaction of multiple TFs with the gene promoter is sufficient to explain the process of transcription. In contrast recent studies have provided results to show that most eukaryotic genes possess multiple TSSs and by extension multiple promoters. These multiple promoters drive gene expression in a context-specific manner [72]. Possession of multiple promoters by extension generates diversity and complexity in the eukaryotic transcriptome. In this study, approximately 1300 core promoters were extracted from the recently assembled *G. morsitans* genome. 65% of tag clusters were identified in the putative 5'UTR and the first coding exon. The classification scheme employed showed that 95% of the core promoters in this dataset are of the broad type. Within the broad core promoter category, 76% do not have preference for one initiation site. These are known as “broad without peak promoters” in contrast to the 24% broad with peak promoters. Narrow core promoters that initiate over tens of nucleotides, constitute a very small proportion (6%) of the total promoter count.

This distribution is concordant with what is emerging regarding metazoan transcriptional programs whereby few of the core promoters fit the “traditional” model of transcriptional regulation, that is, the narrow category. Majority of metazoan genes' core promoters are of the broad type [53]. Broader TSS initiation patterns may in theory be a consequence of non-specificity in the basal transcription machinery, and biological effects of such alterations on transcription yet to be elucidated [80]. With the data utilized in this study, we could not determine whether TSSs in the *G.morsitans* genome are overall defined by only these patterns or if additional data would lead to other patterns.

### ***G.morsitans* core promoters exhibit a propensity for AT dinucleotides**

The nucleotide frequency and distribution in *G.morsitans* core promoters is similar to that of *D.melanogaster* [70] where the core promoters are characterized by propensity for the AT dinucleotides. In *D.melanogaster*, AT enrichment peaks at approximately -200 bp from the TSS. Microarray analysis showed these regions as nucleosome free and mostly for active genes in *D. melanogaster* [254] and *S.Cerevisae* [255]. The positioning of nucleosomes along chromatin has been associated with eukaryotic gene expression regulation because packaging of DNA into nucleosomes affects sequence accessibility. On the other hand mammalian promoters exhibit propensity for the CG dinucleotides. Generally, vertebrate promoters have been associated with presence of CpG island, for instance, in the human genome, half of protein coding genes harbor CpG islands [256] [257]. The difference in dinucleotide preference suggests a fundamental difference in global promoter architecture between mammals and insects. Perhaps, other mechanisms may perform the role of CpG islands in *G.morsitans* and *D.melanogaster*. These mechanisms have yet to be elucidated. While profiling ascidian promoters, Okamura and colleagues [95] postulated that CpG islands are not ancient enough to be found in invertebrates and that these islands may have arose early in vertebrate evolution via some active mechanism. The islands may have since been retained as part of vertebrate promoters. Indeed, introducing an artificial CpG island into mouse cells led to establishment of epigenetic patterns typical of promoters suggesting that mammalian CpG islands might be primed to be promoters by default [53] [89] [258] .

### **Known core promoter motifs are present in *G. morsitans***

To further validate the reliability of our TSSs identification method, the presence of canonical core promoter motifs was examined. Variations in motif frequencies in narrow and broad promoters were investigated. Narrow promoters are characterized by only one or a few consecutive TSSs and are associated with genes that are ex-

pressed in tissue-specific manner. These promoters are enriched for the TATA-box motif. Broad promoters contain several TSSs over a large genomic window (usually not greater than 100 bp). In mammals, they are CpG rich and are usually found in constitutively expressed genes, see review by Sandelin *et al.*, [89]. In this study, high frequency of the TATA-box, INR and DPE motifs was observed in narrow promoters. A similar pattern is observed for the broad with peak category. Since both the narrow and broad with peak classifications harbor a single dominant peak, these core promoter patterns may indicate the specificity of the transcription initiation machinery for peaked promoters. The BREd motif recorded highest frequency in broad without peak promoters indicating that it may be frequently utilized to anchor basal transcription machinery for promoters with multiple TSS.

Despite its conservation in all eukaryotes, comprehensive analyses of *Drosophila* core promoters as well as mammals have suggested that the TATA-box occurs in approximately 10–30% of all genes within a genome [67] [77] [213] [214] [260]. In this study, 23% of *G.morsitans* core promoters harbored a TATA-box. Apart from the BREu motif, TATA-Less promoters record a higher frequency of all other core promoter motifs. A similar observation was made by Gershenzon and colleagues [234] where they postulate that other core promoter motifs may provide a binding site for the basal transcription machinery in the absence of a TATA-box to mediate transcription. Indeed, the DPE was discovered through the analysis of the binding of purified TFIID to TATA-less genes [75].

### **Motif co-occurrence frequencies vary across different core promoter classes**

Most core promoters have at least one core-promoter motif at a functional position working as anchors for the basal transcription initiation machinery. However, the presence of a synergetic combination of two core promoter element is often considerably stronger than a single element as it dictates the position of the TSS. It is extremely rare for all motifs to be present in any given core promoter. Analysis of *G.morsitans*

core promoters two-way motif co-occurrence revealed that the TATA-INR pair has the highest frequency among narrow core promoters whilst the MTE-DPE pair has the highest frequency for broad core promoters. Since high frequency of co-occurrence may indicate that the motifs exert their functions co-operatively, we postulate that the corresponding TFs for the TATA and INR as well as MTE and DPE exhibit synergistic interactions during transcription initiation for narrow and broad core promoters respectively. Indeed, the TATA-INR and MTE-DPE co-operation has been reported by other studies such as [230] [261].

In the broad without peak category, the highest frequency of three-way motif co-occurrence was found to be the INR-MTE-DPE triplet. Some studies on core promoter motifs have shown that neither the DPE nor the MTE exhibits core promoter activity in the absence of an INR [44]. Furthermore, our analysis of TATA-less core promoters shows propensity for the INR and MTE motifs and they have also been shown to compensate for the lack of a TATA-box [42] [65]. Intriguingly, this triplet has anchor points downstream of the TSS, and within the TSS itself. Thus, from a structural point of view this combination may mediate basal transcription initiation without necessarily positioning the RNA polymerase II complex very efficiently. The broad with peak category has the TATA-INR-DPE and combinations as the most frequent triplets. The TATA-INR-MTE combination was also observed by Lim and colleagues [48]. Structurally, these triplets have anchor points on both sides of the TSS and within the TSS itself and may therefore position the RNA polymerase II complex efficiently. TATA-MTE-DPE combination was most frequent in the narrow core promoter category. The MTE exhibits synergy with the TATA and DPE motifs according to Gershon and colleagues [79].

Notably, 26% of the core promoters lack known core promoter motifs, an observation that has been made in other studies [42, 79, 262, 263]. It is hypothesized that undiscovered core promoter motifs may exist. However, the current ones are deemed sufficient to explain the RNA pol II mediated basal transcription initiation program

for majority of genes.

### Several promoters are entangled with repeat sequences

Approximately one third of the *G.morsitans* genome is riddled with repeat motifs. Transposons were contained in 31 out of 1424 core promoters. Recent work by Faulkner and colleagues [124] discovered that retrotransposons and repeat elements are recruited as promoters and there is growing interest on the role of repeat elements in gene regulation. Indeed laboratory investigations have shown specific examples of mammalian genes whose promoters are donated by endogenous transposable motifs. For example, while using reporter constructs for Ewing cell lines, Guillon and colleagues [264] showed that the number of repeats included in the construct highly influenced transcription activation. They postulated that microsatellites in promoters contribute to long-distance transcription regulation. In their review of metazoan promoters, Lenhard and colleagues [53] attribute nearly 200,000 human retrotransposons-driven TSSs identified by CAGE tags. In addition they state that these repeat driven promoters do not so far fit clearly into one of the main promoter classes namely, narrow and broad. According to Cohen and colleagues [265], repeat-recruited promoters have preference for tissue specific activity. A recent study by Lee and Maheshri [266] has shown the indirect impact on gene expression if the repetitive regions contain TFBSs which include transcription factor sequestration, aberrant activation of genes outside given promoter contexts and negative cooperativity in transcription factors. Such occurrences culminate in qualitative changes in the behavior of gene regulatory networks in which target genes are embedded. Vences and colleagues [128], showed that in *Saccharomyces cerevisiae*, as many as 25% of all gene promoters contain tandem repeat sequences. These genes driven by repeat-containing promoters show much higher rates of transcriptional divergence where variations in repeat length result in changes in expression and local nucleosome positioning. This observation could be used in follow-up studies towards understanding of the effect of tandem repeats on transcription control in newly sequenced *Glossina* genomes.

## Intergenic tag clusters may represent non-coding RNA TSS

Non-coding RNA genes consist of abundant and functionally important RNAs comprising several groups involved in distinct cellular processes. Out of 3134 clusters fitting our inclusion criteria, 87 were located on gene-less scaffolds, whilst 1014 were located outside the candidate genic regions. These tag clusters constitute approximately one third of the total count. We refer to them as intergenic tag clusters. However, we are aware that the term ‘intergenic’ is loosely defined as the genome is yet to be fully assembled. In addition manual refinement of the predicted gene models is on-going. In addition manual refinement of the predicted gene models is on-going. In insects, non-coding RNAs appear to occur primarily in intergenic and intronic sequences and at intron-exon junctions. In addition they are significantly associated with genes encoding developmental regulators [267] [268]. We postulate that these intergenic tag clusters may represent TSSs for several classes of non-coding RNA genes in *G.morsitans*. Exploration of intergenic regions with TSS tag clusters may cast new insights into the role of non-coding genomic regions in *Glossina* spp evolution.

## *G.morsitans* promoters are characterised frequently occurring genes in development

Given that TSS-seq has demonstrated to be successful in collecting precise information on TSSs together with digital expression profiling, it was anticipated that tag clusters with many TSS-seq reads are preferentially expressed during the larval and pupal developmental stages of *G.morsitans*. GO analyses was done to assess whether these genes were associated with specific developmental functions categories based on the three core promoter classifications. Genes involved in chitin metabolism were over-represented in all core promoter classes. This underscores their importance during development. Chitin metabolism is crucial to insect morphogenesis which primarily



relies on the ability to remodel chitin-containing structures [253]. The presence of chitin metabolism genes in all promoter classes suggests members of this gene family in *G.morsitans* are transcribed using both narrow and broad programs.

The GO term ‘structural constituent of ribosome’ was over-represented in the broad with peak core promoter category. Genes associated with this ontology term have also been reported in *A.stephensi* embryo transcriptome [269]. The small GTPase binding activity constituted the bulk of GO annotations in the broad without peak category. GTPases are required for several developmental events such as organization of the actin cytoskeleton and signaling by c-Jun N-terminal kinase and p38 kinase cascades [270] [271]. They have also been shown to participate in dorsal closure of the *Drosophila* embryo [272]. Loss of the *Drosophila* larval GTPase Miro has been implicated in dysfunction of the axonal mitochondrial transport, leading to abnormal subcellular distribution of mitochondria in neurons and muscles [273]. The GTPase Cdc42 has recently been shown as a vital component during *Drosophila* embryonic development [274]. The GO term heat shock binding activity is also over-represented. In *Aedes aegypti* larvae and pupae this protein family has been shown as an important indicator of stress and may function as crucial proteins to protect and improve survival [275]. During embryogenesis in *D.melanogaster*, expression of HSP60A is post-transcriptionally regulated in a highly dynamic order, even under heat-shock conditions suggesting novel roles for HSP60 family proteins throughout *Drosophila* development [276].

Other over-represented GO terms include serine/threonine kinase activity, tyrosine kinase activity transferase activity, phosphorus-containing group’s transferases and oxidoreductase activity and DNA binding. These ontologies represent genes that may be involved in signaling networks facilitate cell fate specification during development.

## 2.7 Conclusion, limitations and future work

Most experimental approaches for TFBSs identification rely on previous predictions using computational frameworks. Though experimental confirmation remains the highest form of TFBSs validation, there are some limitations, for example, some assays may not distinguish functional from non-functional TFBSs. In addition, computational prediction suffers high rates of false positive prediction. Due to the shortcomings of both computational and experimental methods for TFBSs identification, effective elucidation of genomic DNA regulatory potential requires collaboration between the two approaches.

In this study, a comprehensive *in silico* analysis of core promoters in the newly sequenced *G.morsitans* genome was done as a starting point to expedite experimental studies. By locating *G.morsitans* TSS using experimental data, the study has provided insight into the promoter architecture of *G.morsitans*. Different initiation patterns were linked to distinct core promoter motifs. To our knowledge, this is the first study to locate TSSs and core promoters in the newly sequenced *G.morsitans* genome. A total of approximately 1300 genes harboring a strong transcriptional signal were obtained and experimentally derived position weight matrices were used to model canonical core promoter motifs. Twenty-six percent of core promoters did not harbor the canonical core promoter motifs, postulating the existence of some undefined mechanisms of transcription control.

Results presented herein have generated testable hypothesis. Validation by experimental methods for some of these predictions would facilitate in assessing the reliability of our computational predictions. The observation that, not all *G.morsitans* promoters harbor canonical core promoter motifs leads to the question of what other mechanisms may facilitate control of the *G.morsitans* basal transcription initiation programs. In humans, core promoters lacking canonical motifs were shown to utilise upstream enhancers to recruit the pre-initiation complex. The presence of repeat

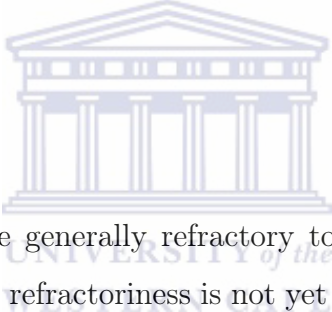
elements in the milieu of the core promoters warrants further investigation into characteristics of repeat-recruited promoters to establish their features and mechanism of action. Tag clusters mapping outside candidate regions could represent non-coding RNA start sites. Characterization of these sites using both computational and experimental frameworks would be essential in elucidating the operation of non-coding RNA in *G.morsitans*. This study provided useful insights using *G.morsitans* genome data that would be employed as a platform to assess the conservation of transcriptional control mechanisms across the yet to be sequenced *Glossina* genomes.

One of the limitations of the data presented in this study is that TSS-seq libraries were sampled from only two developmental tissues. As such, the study does not represent the complete repertoire of promoters in operation in the *G.morsitans* genome. Further, approximately 10 million TSS-seq reads were discarded during the stringent quality control procedure and a conservative approach was employed while defining the minimum number of tag clusters for a significant transcriptional signal. Sequencing of additional libraries in the future will provide additional information, specifically regarding unique aspects of Tsetse biology such as reproduction. Unlike other insects, Tsetse females reproduce by adenotrophic viviparity. Accordingly Tsetse reproduction may exhibit a unique transcriptional program relative to other dipterans and such analysis may further give insights into the evolution of reproductive biology. Additional libraries covering mechanisms that are crucial for Tsetse-Trypanosome interactions such as immune responses, visual, olfactory and salivary components would facilitate understanding of transcriptional control networks that facilitate these mechanisms. Additionally, ChIP-seq and nucleosome-seq assays would be used to facilitate location of promoters of weakly expressed genes. Incorporation of such assays with aforementioned libraries would provide a platform for integrative interpretation of the data. Ultimately, this would provide valuable information that can be used to design rapid and detailed functional assays to attain a more comprehensive understanding of transcriptional programs in different *Glossina* genomes.

# Chapter 3

## *In silico* analysis of promoters of *Glossina morsitans* immunity genes.

### Abstract



**Background:** Tsetse flies are generally refractory to Trypanosome infection. Although the molecular basis for refractoriness is not yet completely understood, it has largely been attributed to a robust immune system. The immune system is facilitated by a series of pathways that culminate in release of anti-pathogen molecules. Availability of macromolecules that facilitate the immune response in a timely and tissue-specific manner is fundamentally controlled at the transcription level. Computational predictions are useful for directing experimental resources to regions most likely to exhibit a biological function. As such, promoter profiling of *Glossina morsitans* immunity genes would be an important step towards understanding how the immune response is coordinated. This is a crucial step towards elucidating the biological complexity of vector-parasite interactions. In this study, transcription factor binding sites for proximal promoters of *Glossina morsitans* immunity genes were characterized as an extension of the previous chapter that dealt with core promoters.

**Methodology:** A comparative genomics approach was employed to identify *Glossina morsitans* immunity genes using orthologous genes in *Drosophila melanogaster* and

select blood-feeding insects. The pipeline for transcription start site (TSS) elucidation outlined in chapter two was used to identify TSS for immunity genes. Proximal promoter regions of the immunity genes were extracted after which an *ab initio* methodology was devised to build transcription factor binding site profiles. Transcription factor binding site (TFBSs) profiles were compared with experimentally determined motifs from the JASPAR insect and vertebrate databases. Those TFBSs falling within the 75th percentile were deemed as over-represented and experimental validation implicating them as regulators of immunity pathways was sought from literature.

**Results:** A total of 190 immunity gene families were obtained from the *Glossina morsitans* proteome of which 61 had an experimentally verified TSS. Comparative analysis showed that most of *Glossina morsitans* immunity gene families were systematically reduced relative to other dipterans. Highly expressed genes code for developmental and immunity programs such as autophagy and apoptosis. The Homeo-box class of transcription factors constituted majority of transcription factors identified as putative immune regulators of *Glossina morsitans* immunity. Overrepresented TFBSs were found to be implicated in control of not only immunity but also development transcription programs.

**Conclusion:** The *ab initio* methodology employed in this study facilitated the identification of TFBSs that had experimental evidence as immune regulators. The study also demonstrated that promoters of apoptosis and autophagy genes that are used for different physiological processes such as immunity and development harbour TFBSs implicated in both processes. The analysis has generated *in silico* inferred hypotheses that can be used to generate regulatory networks or be tested experimentally in subsequent studies.

## 3.1 Insect immunity: genes and pathways

Despite being efficient vectors for disease transmission, the prevalence of Tsetse infection in the field is surprisingly minimal and this phenomenon is referred to as refractoriness. A study conducted by Lehane and colleagues [14] showed that under ideal laboratory conditions, 40% of Tsetse flies fed on *Trypanosoma brucei* were refractory while 90% of remainder flies self-cured from the third blood meal onwards. The inherent mechanisms facilitating self-curing are yet to be established but have been largely attributed to a robust immune arsenal that the fly releases to counter invading pathogens [12,13,15]. Insect immune responses encompass a multi-layered system operating at different levels: (i) physical barriers such as the cuticle and peritrophic matrix, (ii) cellular defences that consist of protease cascades which invoke phagocytosis by haemocytes and melanocytes, and (iii) humoral defences that produce and release reactive oxygen intermediates as well as antimicrobial peptides [277]. Cellular and humoral defences constitute the pathogen surveillance and elimination mechanisms. They are composed of a series of signaling pathways which are activated via three distinct phases; pathogen recognition, signal transduction and pathogen elimination.

### 3.1.1 Pathogen recognition

Upon infection, molecules of microbial/parasitic origin are recognized as ‘non-self’ by pattern recognition receptors (PRRs) which bind to the pathogen associated molecular patterns (PAMPs) including lipopolysaccharides (LPS), peptidoglycans (PGN) and beta-1,3-glucans [277–282]. Some of the well-studied invertebrate PRRs are the peptidoglycan recognition proteins (PGRPs) and the Gram-negative bacteria-binding proteins (GNBPs) [282–287]. Other PRRs include the thioester-containing proteins (TEPs), leucine-rich immune proteins (LRIMs) and C-type lectins (CTLs). In *G.morsitans*, PGRP-LB was shown to down regulate the immune reaction against bacterial infection [288] and as an environment modulator to allow for coexistence

between *G.morsitans* and its symbionts [289].

### 3.1.2 Signal transduction and modulation

Pathogen recognition is followed by transmission of the signal originating from PAMP-associated PRRs to the effector genes via signaling pathways that encompass a series of proteolytic cascades. The insect immune signal transduction is characterized by the classical pathways that include the Toll and IMD pathways (Figure 3.1). While the IMD pathway is activated in response to gram negative bacterial infection, the toll pathway is activated in response to fungi and gram positive bacterial infection. The IMD pathway is primarily involved in the regulation of epithelial immune responses [290]. Both the toll and IMD pathways culminate in the activation and nuclear translocation of NF-, TFs Dorsal/Diff (Toll pathway) or Relish (IMD pathway) prompting transcription of antimicrobial peptides (AMPs) [290].

The prophenol-oxidase (PPO)/melanization cascade in insects is a unique defense mechanism. This cascade initiates with the enzymatic processing of inactive prophenol-oxidases to active phenol-oxidases (POs). Active POs subsequently polymerize to melanin [290].

The JAK/STAT pathway has been associated with immune defense against pathogenic bacterial [291] and viral infections in *D.melanogaster* [292]. The JAK/STAT signaling is relatively simple, with only a few principal components whose activation is commonly associated with stress and cellular damage due to infection. Components of the JAK/STAT pathway include the receptor domeless (Dome), the kinase Hopscotch (Hop), and the TF STAT92E (STAT) (see figure 3.1).

### 3.1.3 Pathogen elimination

Each of the aforementioned signalling pathways culminates in the production of immune effectors. Immune effectors may be one of circulating AMPs (humoral) or phagocytosis, encapsulation and formation of melanotic clusters (cellular).

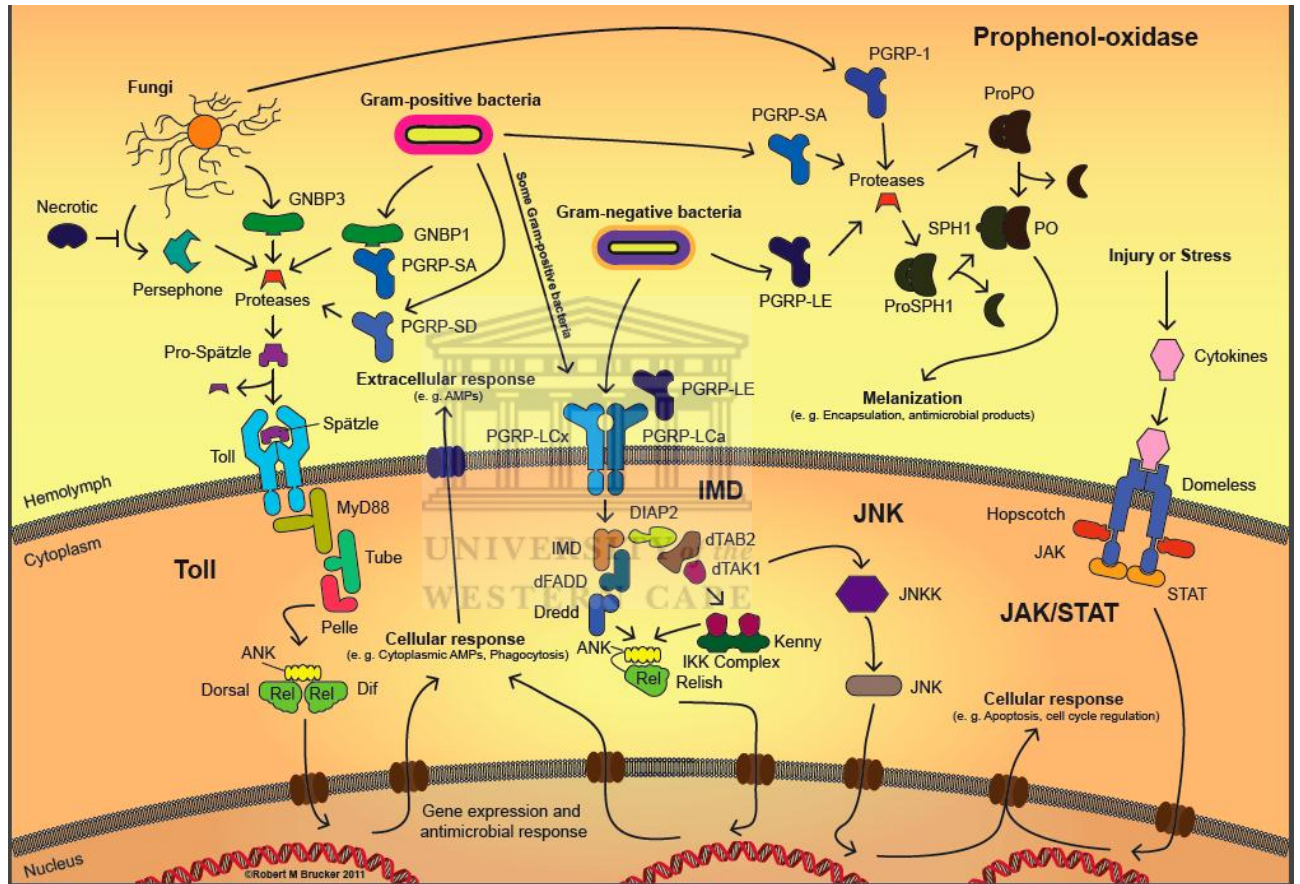


Figure 3.1: Generalized insect innate immune pathways based on *Drosophila* literature. The figure depicts the insect immunity pathways and the genes involved in eliciting an immune response [293].

Several classes of AMPs exhibit wide and complementary spectra of activity against various microorganisms. In *G.morsitans*, some AMPs including Attacin, Cecropin, Defensin and Diptericin have been characterised [12, 13, 15] [294] [295].



During phagocytosis, haemocytes engulf target pathogens as well as apoptotic bodies. Phagocytosis in *Drosophila* is ordinarily mediated by PRRs, for example, PGRP-LC is involved in the phagocytosis of gram-negative bacteria [286]. Thio-ester containing proteins (TEPs) have been proposed to function either as opsonins that promote phagocytosis (like complement C3) or as protease inhibitor (like alpha-2-macroglobulin) in insects [296].

During cellular encapsulation insect lamellocytes form a multilayered capsule around large invaders such as parasitoids in the haemocoel. This results in their isolation, immobilization and subsequent killing by asphyxiation, oxidation or melanization [297]. The melanization cascade effect is two-fold with regard to immunity. On the one hand, intermediates of this cascade generate toxic reactive oxygen species (ROS) that together with melanin are thought to combat infection. Secondly, in *Drosophila*, polymerised melanin contributes not only to wound healing but also to encapsulation of foreign objects, such as parasitoid eggs [298] [299].

Reactive oxygen species are a major component of insect immunity [290] [299] [300]. In *G.morsitans*, increased hydrogen peroxide and nitric oxide levels are induced in the proventriculus following a trypanosome challenge [294]. In addition, genes involved in oxidative stress are induced in midgut transcriptome of infected and self-cured flies citelethane [301]. A recent study comprehensively described the effect of nitric oxide synthase, a Duox and oxidation resistance 1(OXR1) genes in the immune responses of *G.morsitans* [302].

## 3.2 Characterization of promoters of insect immunity genes

Most of the promoter characterization in the insect vectors *A. gambiae*, *A. aegypti* and *C. quinquefasciatus* has hitherto been done using experimental approaches to analyze the promoter sequences of immunity gene families. For example, initial comparative analyses of *A. gambiae* Defensin 1 and two isoforms of *A. aegypti* Defensins identified key regulatory elements responsible for the temporal control of mosquito Defensin gene expression. Defensin 1 promoters of *A. gambiae* and *A. aegypti* were shown to be up-regulated upon immune challenge [303] and this stimulated activity was shown to depend upon a cluster of three NF-kappaB TFBSs and closely associated C/EBP-like motifs, which function as a unit for optimal promoter activity. KappaB-like motifs were abundant within AMP gene promoters and most are very closely associated with putative CEBP binding sites. The study concluded that novel association between NF-kappaB and CEBP binding sites might, therefore, be of broad significance.

By employing microarray data to analyse motifs that were over-represented in the 5'UTRs of up-regulated genes, Hernandez and colleagues [304] were able to identify experimentally verified immune-related TFBSs. In addition, this study demonstrated that immunity and related genes in *A. gambiae*, *A. aegypti* and *D. melanogaster* share enrichment of A-T rich motifs. In another study Sieglaff and colleagues [305] employed comparative genomics to locate regulatory elements in regions flanking the 5' UTR of orthologous genes in *A.aegypti*, *A. gambiae* and *C.quinquefasciatus*. These analyses identified several motifs representing 18 families of putative cis-regulatory elements conserved among the three mosquito species relative to *D. melanogaster*. Some of the motifs had been experimentally verified as TFBSs.

With the availability of the *G.morsitans* genome data, analysis of TFBSs profiles in immunity genes would provide a crucial link between the genome and dynamic aspects of gene expression and regulation.

### 3.3 Promoters of immunity genes as a potential tool for design of novel vector control methods

Genetic methods of vector control mostly rely on effector genes and/or toxins that either kill the parasites or interfere with vector-parasite interactions. Among the key mechanisms of vector-parasite interactions is mounting of an immune response because it facilitates parasite elimination. Availability of macromolecules that facilitate the immune response in a timely and tissue-specific manner is fundamentally controlled at the transcription level. As such, identifying mechanisms responsible for transcriptional control of such processes/pathways is a crucial step towards elucidating the biological complexity of vector-parasite interactions. Successful investigation of TFBSs within promoters would require an integrated approach that exploits both computational and experimental techniques. Computational predictions are useful for directing experimental resources to regions most likely to exhibit a biological function. As such, promoter profiling of *G.morsitans* immunity genes would be an important step towards understanding how the immune response is coordinated at transcriptional level and to lay a foundation for targeted experimental studies.

Similar to chapter two, this chapter relies on the availability of *G.morsitans* genome allowing interrogation of the organization of TFBSs modules in gene promoters using *in silico* methods. However, while chapter 2 made a global examination of core promoters in *G.morsitans* genome, the analyses in this chapter focuses on proximal promoters of immunity genes.

### 3.4 A summary of chapter objectives

The overall aim of this chapter was to characterize the promoters of *Glossina morsitans* immunity genes. The objectives are outlined below:

- 1) To identify immunity genes in the *G.morsitans* genome based on orthologous relationship with other dipteran species.
- 2) To identify immunity genes with experimentally verified TSS based on the mapping profile obtained in chapter two.
- 3) To locate TFBSs on promoters of immunity genes using *de novo* approaches and assess TFBSs overrepresentation.



## 3.5 Methodology

### 3.5.1 Compilation of insect proteomes and immunity genes

The IMMUNODB database [306] is a compilation of insect immune and related gene family assignments together with their phylogenetic data. Using the menu for viewing expert annotations, protein sequences from twenty-seven immune and related gene families from select insect vectors *A.gambiae*, *A.aegypti* and *C.quinquefasciatus* were obtained. Additionally *D.melanogaster*'s immune gene set was included as it is arguably one of the best characterized dipteran genomes. Immunity gene's external IDs were catalogued for use in obtaining corresponding orthologous genes in *G.morsitans* via BIOMART [307]. A full complement of protein coding genes for the abovementioned species was downloaded using the BIOMART tool of the ENSEMBL metazoa database [308]. *G.morsitans* full protein complement constitutes 12,220 genes which is somewhat comparable to that of *A.gambiae* with 12,810 genes (Table 3.1).

Table 3.1: A summary the total number of protein coding genes for selected insect vectors

Name of insect	Number of protein coding genes
<i>Aedes aegypti</i>	15,998
<i>Anopheles gambiae</i>	12,810
<i>Culex quinquefasciatus</i>	18,955
<i>Drosophila Melanogaster</i>	13,937
<i>Glossina morsitans</i>	12,220

To establish orthologous relationships, ORTHOMCL [309] was run with the default parameters. Briefly the ORTHOMCL algorithm performs an all-against-all alignment using BLAST [310] and finds reciprocal best similarity pairs between species using putative orthologues. The similarity matrix is normalized by species after which markov clustering is applied to identify orthologues groups as well as paralogues. Re-

sults generated by ORTHOMCL were used to extract clusters of orthologous groups (COGs) and importantly to identify immunity genes based on external IDs compiled from IMMUNODB.

### **3.5.2 Identification of immunity genes with a transcriptional signal**

The TSS elucidation pipeline employed in chapter two was used for TSS location and thereby promoter identification. A tag cluster with at least 10 TSS-seq tags was deemed a sufficient transcriptional signal and was used to extract all immunity genes with a sufficient transcriptional signal.

### **3.5.3 Identification of immunity genes with possible developmental roles**

Given that the TSS-seq data used in this study was sampled from developmental stages and TSS-seq is coupled with digital expression profiling, immunity genes with a transcriptional signal were evaluated as possible regulators of developmental processes. Absolute tag counts were obtained for each gene and the summary statistics computed. Genes with greater than or equal to the mean of tag counts for the immunity gene promoters dataset were considered highly expressed. Corresponding gene family annotations for these highly expressed immunity genes were obtained and evaluated for experimental evidence associating them with developmental processes.

### **3.5.4 Promoter extraction**

Several studies have showed that 80% of functional TFBSs are located within approximately 1kb upstream of the TSS [311–314] . This analysis was performed using

-1000/+100 region surrounding the TSS. To avoid confounding results, promoter sequences whose composition constituted at least 2% repeat sequences were excluded from further analysis.

### 3.5.5 Identification of transcription factor binding sites

An *ab initio* motif discovery strategy was employed to identify TFBSs in *G.morsitans* immunity gene promoters. Tompa and colleagues [195], showed that even though numerous *ab initio* motif discovery programs have been developed, none of them shows a distinct advantage over the others on all data types. IMPROBIZER [315], MEME [316] and INFOGIBBS [317] were selected for motif discovery. The reason for selecting these algorithms was to compensate for possible deficiency of the search algorithms such that a motif identified by at least two of these prediction programs would be deemed as high ranking. The underlying algorithm for IMPROBIZER and MEME is expectation maximization [189] while INFOGIBBS employs Gibbs sampling [175]. For each promoter set, the parameters for motif searching were defined according to specifications of the search algorithm. For example, IMPROBIZER allows for a maximum of six motifs per run, while MEME is flexible with regard to the number and width of motifs. On the other hand INFOGIBBS allows a specific width but variable number of motifs per run.

For IMPROBIZER, six motifs was generated in each of the promoter sets, while for MEME ten motifs using the zoops model were generated. For INFOGIBBS, several iterations were performed to generate motifs whose widths ranged from 7 to 15 nucleotides. PWMs identified by at least two of the search algorithms were considered highly significant. In addition those that were not found by at least two search algorithms but had very low p values were also combined with the highly significant set.

*Ab initio* searches were compared with experimentally verified PWMs from JASPAR insect and vertebrate PWMs [244] using the STAMP [318] tool. The alignment with

the lowest E value for each of the PWM was deemed as the best match. The process was repeated for each gene family promoter set and the best matches were compiled.

### **3.5.6 Analysis of TFBSs overrepresentation**

Individual counts of TFBSs were computed. TFBSs falling within the 75th percentile were considered overrepresented and the abundance of corresponding TF family was obtained .





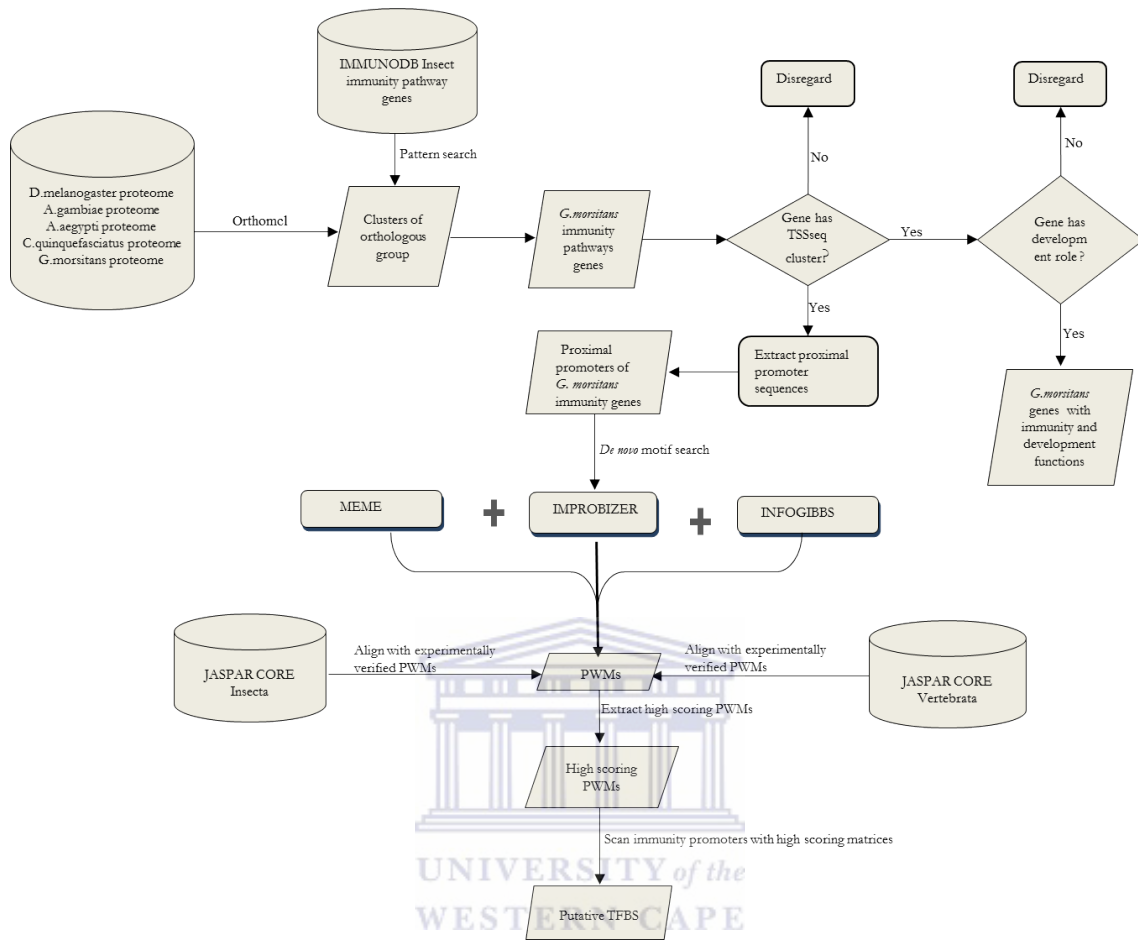


Figure 3.2: Protocols and tools used for promoters of immunity gene's promoters analysis.

## 3.6 Results

### 3.6.1 Comparison of the number of immunity genes between *G. morsitans* and other dipterans

ORTHOMCL clustering of the insect proteomes generated 16,452 clusters. These clusters included 190 immunity gene families. A comparison of immunity gene family numbers with select insect vectors is shown in Table 3.2 (for a graphical summary, see appendix four). The *G.morsitans* genome encodes genes for all immunity pathways. Apart from autophagy genes, galactoside binding lectins, Toll- like receptors and Toll pathway members, there is a systematic reduction in *G.morsitans* immunity gene counts relative to other insects (Table 3.2). The JAK/STAT pathway members and small RNA regulatory pathway members have higher gene counts compared to *D.melanogaster*.

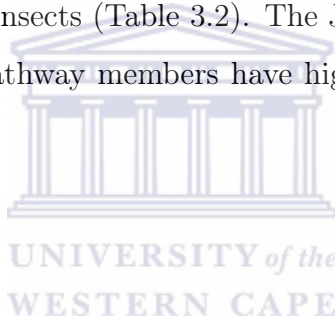


Table 3.2: A summary of the whole complement of immunity genes in selected insect vectors

Immunity gene family	<i>G.morsitans</i>	<i>D.melanogaster</i>	<i>A.aegypti</i>	<i>A.gambiae</i>	<i>C.quinquefasciatus</i>
Antimicrobial peptides (AMPS)	7	20	18	19	17
Autophagy genes (APHAGS)	14	16	19	19	19
1,3D glucan binding proteins (BGBPS)	2	3	7	7	11
Caspase activators (CASPAS)	6	7	3	2	3
Caspases (CASPS)	3	4	11	14	16
Catalases (CATS)	1	2	2	1	1
CLIP-domain serine proteases (CLIPS)	30	50	62	55	78
C-type lectins (CTLS)	5	34	40	25	54
Fibrinogen related proteins (FREPS)	3	13	37	52	77
Galactoside binding lectins (GALES)	6	6	12	10	11
Inhibitors of apoptosis (IAPS)	3	4	5	8	6
IMD pathway members (IMDPATHS)	4	8	5	7	8
JAK/STAT pathway members (JAKSTATS)	5	3	3	4	5
Lysozyme (LYS)	2	13	7	8	4
MD2-like receptors (MLS)	5	10	26	15	19



Immunity gene family	<i>G.morsitans</i>	<i>D.melanogaster</i>	<i>A. aegypti</i>	<i>A.gambiae</i>	<i>C. quinquefasciatus</i>
Peptidoglycan recognition proteins (PGRPS)	4	12	8	11	10
Peroxidases (PRDXS)	12	19	20	26	19
Prophenol-oxidases (PPOS)	2	3	10	9	9
Relish-like proteins (RELS)	1	3	3	2	3
Scavenger receptors(SCRS)	15	23	20	19	22
Superoxide dismutases (SODS)	2	4	6	5	5
Speatzle-like proteins (SPZS)	4	6	9	6	7
Serpins (SRPNS)	14	29	29	21	40
Small RNA regulatory pathway members (SRRPS)	22	14	32	20	37
Thio-ester containing proteins (TEPS)	4	6	8	13	10
Toll-like receptors (TOLLS)	9	9	11	10	9
Toll pathway members (TOLLPATHS)	5	5	5	5	5

Immunity gene counts on all but *G.morsitans* are based on data from the original version of the ImmunoDB resource [306].

## 3.6.2 Promoter extraction and TFBSs analysis

### 3.6.2.1 Tag cluster distribution

TSS-seq data was mapped to the *G.morsitans* genome and TSS-seq clusters delineated as described in chapter two. Sixty-one *G.morsitans* immunity genes had an experimentally determined TSS. Genes implicated in control of developmental programs had higher expression levels as their corresponding tag clusters had greater than 300 (mean number of tags for this dataset) reads (Table 3.3 highlighted in green). They include autophagy genes, caspases, and inhibitors of apoptosis, scavenger receptors and serine protease inhibitors (serpins).



Table 3.3: Summary of immunity genes with experimentally verified TSS

Immunity gene family	Total number of genes	Genes with TSS tags	Total Number of TSS tags
Antimicrobial peptides	7	3	77
<b>Autophagy genes</b>	<b>14</b>	<b>5</b>	<b>873</b>
1,3D glucan binding proteins	2	0	0
Caspase activators	3	2	33
<b>Caspases</b>	<b>6</b>	<b>3</b>	<b>346</b>
Catalases	1	1	61
CLIP-domain serine proteases	30	6	191
C-type lectins	5	0	0
Fibrinogen related proteins	3	1	84
Galactoside binding lectins	6	2	22
<b>Inhibitors of apoptosis</b>	<b>3</b>	<b>2</b>	<b>541</b>
IMD pathway members	4	3	83
JAK/STAT pathway members	5	1	57
Lysozyme	2	0	0
<b>MD2-like receptors</b>	<b>5</b>	<b>3</b>	<b>1649</b>

Immunity gene family	Total number of genes	Genes with TSS tags	Total Number of TSS tags
Peptidoglycan recognition proteins	4	1	170
Peroxidases	12	2	118
Prophenol-oxidases	2	2	23
Scavenger receptors	15	3	307
Superoxide dismutases	2	0	0
Speatzle-like proteins	4	1	10
Serpins	14	6	965
Small RNA regulatory pathway members	22	10	707
Thio-ester containing proteins	4	0	0
Toll-like receptors	9	0	0
Toll pathway members	5	4	203
Total	190	61	6520

### 3.6.2.2 PWM distribution

Half of putative TFBSs on *G.morsitans* immunity promoters matched *D.melanogaster* PWMs (Figure 3.3).

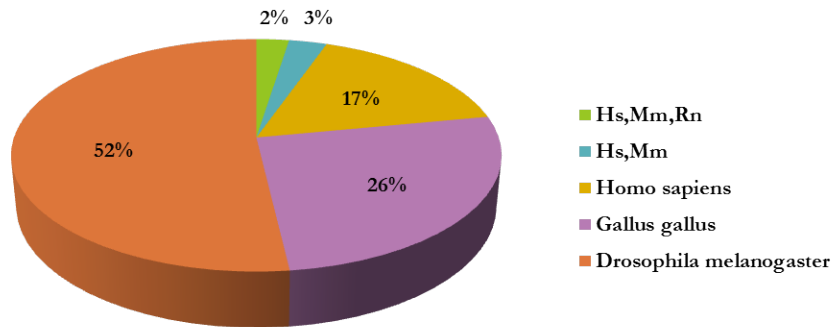


Figure 3.3: Distribution of organism PWM instances that are over-identified in *G.morsitans* promoters of immunity genes. PWMs originating from *D.melanogaster* constitute half of putative TFBSs. Other vertebrate promoters are abbreviated as follows: Hs: *Homo sapiens*; Mm: *Mus musculus*; Rn: *Rattus norvegicus*.

### 3.6.2.3 Transcription factor family distribution

As shown in Figure 3.4, the homeo-box family of TFs recorded majority of TF families (45%). The homeo-box family of TFs regulates wide-ranging crucial activities during development including directing the formation of limbs and organs along the anterior-posterior axis and regulating cell differentiation [319]. Since the datasets employed for TSS elucidation were sampled from *G.morsitans* larvae and pupae, high frequency of homeo TFBSs is an indication of their important role in *G.morsitans* developmental programs.



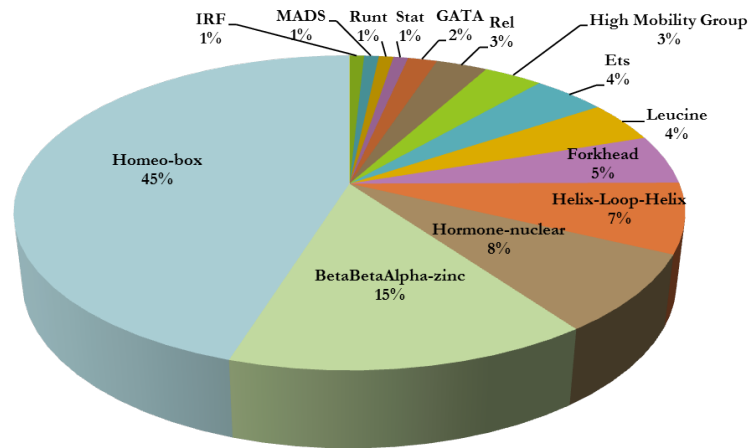


Figure 3.4: Summary of transcription factor family instances. The homeo-box family of transcription factors is highly represented with 45% instances for the whole dataset.

#### 3.6.2.4 Frequently occurring transcription factor PWMs

All TFs occurring within the 75th percentile were considered as frequently occurring. There are forty-three in total and their counts are displayed in Figure 3.5. The most frequently occurring transcription factor PWMs are ZNF354C, IRF1 and Egr1 as they were present in more than half of *G.morsitans* immunity promoters; 33, 36 and 37 respectively.



Figure 3.5: Graphical representation of TF IDs with high frequency. The IRF1, ZN354C and Egr-1 had the highest frequency of occurrence.

### 3.6.3 Predicted and experimentally verified transcription factor binding sites

Several experimentally implicated regulators for some gene families including antimicrobial peptides, autophagy genes, caspases, inhibitors of apoptosis and serpins were found in this study (Table 3.4).

Table 3.4: Predicted transcription factor binding site for *G. morsitans* immunity genes

Immunity gene family	Predicted TFBSs with experimental verification	References	Predicted TFBSs without experimental verification
Antimicrobial peptides	NFKB1, GATA, cad, f2_II, NFATC2	[320–324]	Arnt::Ahr, brk, CTCF, E2F1, ESR1, MEF2A, NR3C1, PBX1, PARG::RXRA, REST, SOX, RUNX1, SPI1 AbdB, Arnt::Ahr, brk, Cf2_II, EBF1, HLF, hth,
Autophagy genes	CEBPA, CTCF, Eip74EF, ESR2, Egr1, EWSR1FLI1, Foxd3, MEF2A, NR4A2, Pou5f1, PPARC, run::Bgb, Sox2, SP1, SRF, TP53	[325–335]	INSM1, IRF1, IRF2, MIZF, Myb, NFATC2, NFIC, NHLH1, oc, opa, Pax5, PaxLAG1, PPARC::RXRA, RORA_1, RREB1, slbo, Tal1::Gata1, TFAP2A, TLX1::NFIC, SPI1, twi, vvl, znf143
1,3D glucan binding proteins	N/A	0	0
Caspase activators	N/A	N/A	Arnt, Arnt::Ahr, brk, E2F1, hkb, IRF1, MIZF, Myf, NFE2L2, opa, Pax2, Pax6, RELA, REST, RREB1, run::Bgb, slbo, SOX10, SPI1, T, TP53, vvl, z, Zfp423

Immunity gene family	Experimentally verified TFBSs	References	Predicted TFBSs without experimental verification
Caspases	CREB1,E2F1,Eip74EF,HLF,Klf4, Lag,NFATC2,SP1,Tal1::Gata1, TP53	[328], [336-339]	achi,btd,Cf2_II,CTCF,Egr1,EWSR1,FLII, FEV,Foxa2,gt,Hand1::Tcf2a,HoxA5, NFIL3,NHLH1,odd,Pax4,PBX1,REST PPARG::RXRA,RREB1,Sox17,Sox2,
Catalases	SP1	[340]	SPIB,Tcfcp2l1,tll,znf143,ZNF354C br_Z4,ESR1,Hand1::Tcf2a,INSM1,Myb, Spz1,usp,wl,YY1,Egr1,RREB1,Pax4,znf143 achi,API1,brk,Egr1,FOXA1,Foxd3,FOXF2, gt,Hand1::Tcf2a,HIF1A::ARNT,Hlft,INSM1, IRF1,Klf4,MEF2A,Myb,MYC::MAX, MZF1_513,NFATC2,odd,opa,REST,RREB1, RXR::RAR_DR5,STAT1,T,Tal1::Gata1, TAL1::TCF3,tll,TLX1::NFIC,Zfx achi,brk,E2F1,EBF1,Egr1,ELK4,ISRI,exd, EWSR1,FLI1,Hand1::Tcf2a,hkb,INSM1, IRF1,Myc,NR1H2::RXRA,NR2F1,NR3C1, Pax4,pax5,Pax6,RELA,RREB1,run::Bgb, RXRA::VDR,SRF,Tal1::Gata1,FAP2A, opa,TLX1::NFIC,Tf1
CLIP-domain serine proteases	Gata1,HNF1B	[341] [342]	
Galactoside binding lectins	TP53	[343]	



UNIVERSITY *of the*  
WESTERN CAPE

Immunity gene family	Experimentally verified TFBSs	References	Predicted TFBSs without experimental verification
Fibrinogen related proteins	N/A		Arnt::Ahr,brk,cad,CREB,E2F1,MIZF,Myb,NFYA,Pax6,Pou5f1,T,ZNF354C brk,br_Z4,CEBPA,Egr1,Esrrb,fkh,IRF2,
Inhibitors of apoptosis	Ddit3::Cebpa,NFKB1,STAT1,SP1	[344–346]	Lag1,MIZF,NFE,NFYA,odd,PPARG,RELA,REST,REB12L1::MafG,SPIB,T,usp TLX1::NFIC,ZNF354C
IMD pathway members	IRF1	[347]	Arnt::Ahr,BRCAl,brk,btd,D,FEV,Foxd3,FOXI1,GABPA,gt,Hand1::Tcfe2a,INSM1,Klf4,MIZF,NFE2L2,NHLH1,odd,opa,PLAG1,PPARG::RXRA,run::Bgb,SOX1,SP1,Spz1,SRF,Tal1::Gata,vnd bcd,BRCAl,brk,btd,CREB1,Ddit3::Cebpa,Eip74EF,EWSR1FLI1,exd,MIZF,Myb,NFIC,NHLH1,Nr2e3,odd,opa,Pax5,REST,run::Bgb,SOX10,Sox17,SPI1,STAT1,TLX1::NFIC,Tf1,ZNF354C
JAK/STAT pathway members	AP1,E2F1,IRF1,SP1	[348]	

Immunity gene family	Experimentally verified TFBSs	References	Predicted TFBSs without experimental verification
MD2-like receptors	N/A		Arnt::Ahr,brk,btd,CG7056,CTCF,Egr1,En1,Esrrb,Evi1,EWSR1FLI1,exd,FOXF2,h,hb,hkb,HLF,HNF1B,IRF1,IRF2,Klf4,MIZF,MYC::MAX,NR4A2,nub,Pax4,Pax5,Pax6,PLAG1,PARG::RXRA,REST,RORA_1,run::Bgb,RXR::RAR_DR5,sd,slbo,SOX9,SP1,Spz1,TBP,TFAP2A,tll,TLX1::NFIC,Ty1,Zfp423,Zfx,ZNF35
Peptidoglycan recognition proteins	N/A		NR4A2, usp
Peroxidases	N/A		brk,CEBPA,E2F1,Egr1,Eip74EF,ELF5,ELK4,ESR1,FOXF2,h,IRF1,Klf4,Lag1,MIZF,Mycn,Myf,NHLH1,nub,opa,PLAG1,Pou5f1,RORA_1,run::Bgb,RXRA::VDR,sd,Sox2,STAT1,Stat3,T,TEAD1,TP53,Zfx
Prophenol-oxidases	AP1, REL, RUNX1	[342]	

Immunity gene family	Experimentally verified TFBSs	References	Predicted TFBSs without experimental verification
Scavenger receptors	AP1	[349]	AbdB, achi, Ar, btd, f2_I, Ddit3::Cebpa, E2F1, Eip74EF, gt, h, Hand1::Tcf2a, hkb, INSM1, kni, Kr, MIZF, Myc, Myf, MZF1_513, NHLH1, opa4, pax5, PPARG::RXRA, RREB1, sna, SPIB, Spz1, T, tll, TLX1::NFIC, usp, ZNF354C brk, Ddit3::Cebpa, Deaf1, Egr1, En1,
Speitzle-like proteins	N/A		Klf4, MEF2A, MIZF, NFE2L2, Pax5, Pax6, REST, Tcfcp2l1, znf143
Serpins	AP1, CEBPA, E2F1, Egr1, hb, IRF1, IRF2, Lhx3, run::Bgb, Tal1::Gata1, TP53	[350-358]	achi, Arnt::Ahr, Ddit3::Cebpa, ELK4, ESRI, EWSR1L1I1, exd, Foxd, hkb, hth, kni, MIZF, Myb, Myf, NFATC2, NFE2L2, Nr2e3nub, odd, RORA_1, RORA_2, RREB1, RXR::RA_DR5, Pax5, ovo, SOX10, Sox2, SOX9, T, TEAD1, Trl, zen, ZNF354C



Immunity gene family	Experimentally verified TFBSs	References	Predicted TFBSs without experimental verification
Small RNA regulatory pathway members	GATA3	[359]	BHL1,brk,btd,btn,cad,C'EBPA,CG15696, CREB1,ct,Dr,EBF1,Egr1,Eip74EF, EWSR1FLI1,FEV,fkh,Gfi,h,hkb,INSM1, IRF1,Klf4,Lag1,MAX,MEF2A,Myb MYC::MAX,Myf,MZF1_513,NFE2L2, NFIC,NFYA,NR4A2,nub,onecut,Pax4, PBX1,Pou5f1,PPARG::RXRA,REST, run::Bgb,slbo,sna,SOX10,Sox2,SOX9, SP1,SPI1,SPIB,SRF,in,TLX1::NFIC,TP53, Trl,z,ZEB1,znf143,ZNF354C brk,btd,cad,EBF1,Egr1,Eip74EF, EWSR1FLI1_fkh,GABPA,Gata1,GATA3, Hltf,Klf4,Lag1,Myf,NFIL3,NHLH1, Nr2e3,NR2F1,Pax2,Pax5,Pax6,Pou5f1, RORA_1,RUNX1,SOX10,SPIB,Spz1,TEAD1, T,TFAP2A,zen,znf143,ZNF354C
Toll pathway members	N/A		

## 3.7 Discussion

The global *G.morsitans* core promoter analysis in chapter two included a set of immunity genes. This chapter expands the promoter analysis of immunity genes by examining their proximal promoters.

### **Most of *G.morsitans* immunity gene families are systematically fewer than other dipterans**

This analysis shows that the fundamental set of components that define the insect immune system is present in *G.morsitans* albeit in smaller numbers relative to other dipterans. However, members of the Toll like receptors and JAK-STAT gene families are present in comparable proportions. Species specific immunity gene expansions have been reported in other insects including *A.gambiae* and *D.melanogaster* [360], and *T. castaneum* [361] but they were not observed in *G.morsitans*. Given that *G.morsitans* is a vector and hence exposed to a myriad of pathogens just as all the other vectors used in this study, we propose that *G.morsitans* may employ other mechanisms to fabricate an elaborate immune response. Interestingly, genes encoding members of the small RNA regulatory pathways have comparable numbers to other insects and higher than *D.melanogaster* and *A.gambiae*. No study has investigated the role of small RNAs in *G.morsitans* hitherto, but they have been implicated as regulators of immune response in other insect vectors such as *A.gambiae* [362], *A.aegypti* [363] and *C.quinquefasciatus* [364]. The suggestion that *G.morsitans* probably employs small interfering RNAs to control parasite invasion warrants further investigation into the small RNAs as mediators of immune response in *Glossina* spp.

### **Genes encoding pathways shared by developmental and immunity programs record significant transcriptional signals**

Simply described as self-eating and self-killing respectively Maiuri and colleagues [365]

described the functional relationship between apoptosis and autophagy as complex since they may be triggered by common upstream signals. In this analysis, it was observed that most of the frequently expressed genes (greater than 300 TSS-seq reads) include those encoding the apoptotic and autophagy processes. They include autophagy genes, caspases, inhibitors of apoptosis and scavenger receptors (Table 3.3 highlighted in bright green). Apoptosis allows precise destruction of cells to preserve tissue architecture and integrity while autophagy facilitates cytoplasmic degradation and recycling of unwanted cells. *Drosophila* larvae undergo large cellular remodeling during metamorphosis to reach tissue maturation. Several structures such as the fat body and the salivary glands have to be degraded to develop into the adult organism [366]. In addition a recent study conducted by Denton and colleagues [367] showed that the elimination of larval midgut cells is also dependent on autophagy.

Other genes with significant expression and related to the apoptotic and autophagy processes include the enzymes catalase and peroxidase. Klichko and colleagues [368] showed that the catalase enzyme is highly expressed during development in *D.melanogaster* and is crucial to regulating free radicals generated by apoptosis and autophagy. *Drosophila* components of the JAK/STAT pathway were initially discovered from studies on embryonic development [369–372] and have since been shown to mediate activation of immune responses (see review by Agaisse and Perrimon [373]). In this study a strong transcriptional signal was recorded for a JAK-STAT pathway member. The JAK-STAT pathway plays a myriad of roles during development in *D.melanogaster* for example, embryonic development, haematopoiesis, sex determination, segmentation, gut and tracheal development as well as development of imaginal discs [374].

Other genes with strong transcriptional signals include those that function in the Toll pathway. The Toll pathway was initially identified in a series of genetic screens for genes involved in early *Drosophila* embryonic development [375] and its components were later implicated in eliciting an immune response [376–378]. Members of

this pathway include genes such as the MD-2 receptors, clip-domain serine proteases, peptidoglycan recognition proteins and serpins [379] [380].

Small RNA regulatory pathway members were also identified as significantly expressed. miRNAs have been implicated as important regulators of the *D.melanogaster* developmental processes (see Flynt and Lai [381] for a comprehensive review). Identification of genes implicated in the small RNA regulatory pathways suggests that the operation of small RNA regulatory pathways during *G.morsitans* development.

### **The Homeo-box transcription factors constitute majority of the identified TFs**

Approximately 50% of the TFBSs identified by this study are bound by TFs belonging to the homeobox family. Homeobox TFs were originally defined as regulators of development and differentiation during embryogenesis but have been shown to have diverse physiological roles including immunity. For example, the TF Caudal (*cad*) that has been implicated in control of *Drosophila* embryogenesis was among the TFs whose TFBS fit our inclusion criterion of over-representation. Cad has also been implicated in immune defenses; Junell and colleagues [382] showed that *D.melanogaster's* Cad is involved in control of constitutive AMP gene expression in a tissue and sex-specific manner. Ryu and colleagues [383] have shown that Cad controls the commensal-gut mutualism by inhibiting nuclear factor kappa B-dependent AMP genes in *Drosophila*. This is crucial for maintaining innate immune homeostasis between commensal-gut flora and hosts in *Drosophila*. Recently, Clayton and colleagues (2013) showed that in *A.gambiae*, Cad is a negative regulator of the IMD pathway [384].

### **Transcription factors with high frequency of occurrence have been implicated in control of immunity and development transcription**

Most of the TFBSs obtained had experimental evidence implicating them as regula-

tors of several biological processes most importantly immunity and development. Of interest are the TFBSs present in at least half of the immunity genes, that is, the ZNF354C, IRF1 and Egr1. The Zinc finger protein 354C (ZNF354C) was shown to regulate embryonic development processes [385]. Interferon (IFN) regulatory factor 1 (IRF-1) was originally identified as TF involved in the regulation of the IFN system [386] and was later reported in the upstream region of several genes involved in cell growth control [387]. Though no orthologue for IRF is found in *Drosophila*, in humans, the key pathways involved in mediating apoptosis are regulated by IRF1, STAT1 and NF-kappaB [388]. IRF TFs are specifically activated by the Toll signaling pathway and participate in the critical processes of antiviral innate immunity [389].

*Drosophila* homologue of human Specificity Protein 1 (Sp1) was first described as a head-specific segmentation gene [390]. Sp1 has since been implicated in *Drosophila* anatomical morphogenesis [391] [392]. In mammals SP1 is involved immunity [393] [394] IL-10 mediated immune responses [395] [396]. The *Drosophila* homologue of P53 (DmP53) has been shown to exert dual roles in cell death and cell differentiation [397] and to facilitate rapid induction of apoptosis conveying resistance to viral infection [398]. GATA factors participate in tissue-specific immune responses in *Drosophila* larvae [321]. The E24 TF has also been implicated in immunity of *Drosophila* larvae [399].

Other TFs that have been implicated in immunity include the nuclear factor kappa-B (NFkB) that has been shown as a positive regulator of AMP genes in *Drosophila* larvae [321]. Regulation of AMP production via NFkB has also been demonstrated in *A. gambiae* [400] and *A. culicifacies* [341]. In *A. aegypti* RUNX TFs regulate PPO gene expression [342]. AP-1 and STAT TFs are the major inhibitors responsible for attenuating NF-kappaB-mediated transcriptional activation during the innate immune response in *Drosophila* to dampen the cytotoxic signals [401].

The TF Brinker (brk) controls the c-jun terminal kinase pathway which in turn trig-

gers apoptosis [402]. Nuclear factor of activated T-cells (NFAT/C2) is one of the repress families of TFs that have been implicated in immunity. NFAT signaling was first identified as mediators of adaptive immunity [403]. The SRY-related HMG-box (SOX) family of TFs in *Drosophila* is involved in the regulation of various events of cell determination/differentiation during development [404]. The *Drosophila* homologue of TLX1, c15 has been reported as one of the cell cycle determinants and has been implicated in *Drosophila* development [405]. Myb has been implicated as a key regulator of the programmed death of neural precursor cells at the posterior wing margin in *Drosophila* embryos [406]. Analysis of expression patterns in *Drosophila* embryos revealed that Bgb interacts with runt to exert their function [407]. The existence of this run::bgb duo in the overrepresented set suggests their concerted role in *G.morsitans* developmental pathways. The *Drosophila* homologue of Myf nautilus (nau) has been implicated in larval somatic muscle development [408]. Ecdysone-induced protein 74EF (Eip74EF) is required for the proper functioning of the larval muscles during early morphogenesis [409] and programmed cell death of larval tissues during *Drosophila* metamorphosis [410]. The TF Buttonhead (btd) is expressed during *Drosophila* segmentation [411]. The T-Box TF Brachyury (T) was reported among the genes expressed during midgut morphogenesis in *Drosophila* embryo [412]. The *Drosophila* orthologue of RREB1, Pebbled (also known as Hindsight), negatively regulates muscle development while promoting neuronal development during embryogenesis [413].

### 3.8 Conclusion, limitations and future work

Accurate identification of transcription factor binding sites remains a major challenge in computational biology. Identification of such binding sites would facilitate the development of gene networks to model interactions that would help unravel important biological pathways. As a starting point this study exploited orthologous relationships to obtain *G.morsitans* immunity genes. TSS-seq data was employed to identify immunity genes with sufficient transcriptional signal. Promoters of these genes were analysed to identify common and significant patterns with putative regulatory potential.

Although the direct application of PWMs to scan sequences is known to suffer from high false-positive prediction rate, the study was able to detect several motifs that are experimentally proven TFBSs in their corresponding promoters. These TFBSs have been shown to function as immune regulators. The study has also demonstrated that promoters of genes encoding pathways used for different biological processes such as immunity and contain TFBSs implicated in both processes. The study was able to provide transcription factor binding information for approximately 25% of the total *G.morsitans* immune complement. In future, deeper sampling would facilitate investigation of most if not all of *G.morsitans* immune genes. In addition, availability of more data may extend this study to promoters of genes participating in other pathways that are important to *Glossina* biology especially Tsetse-Trypanosome interactions.

Ultimately, the study has generated hypotheses that can be tested experimentally in future. In addition to validating computationally predicted TFBSs, functional tests would ascertain whether a given binding event activates or represses transcription. Such measured functional outcomes of TF binding would have direct implications for biological networks and would ultimately help examining transcriptional activity in relation to different stimuli. This would expand current understanding of *Glossina*

spp biology and may well facilitate prioritization of candidate genes for further investigation into the mechanisms of Tsetse-trypanosome interactions in an effort to develop novel vector control strategies.





# Chapter 4

## GmPromDB: A database of *Glossina morsitans* promoters.

### Abstract



**Background:** The deluge of biological data ensuing from genome sequencing efforts in the recent past has necessitated development of tools and methodologies for storing and accessing these data sets in a meaningful way for efficient utilization. Databases facilitate storage and accessibility of this data.

**Methods:** A three tier architecture was implemented where a MYSQL relational database was used to store promoter and corresponding gene information. PERL CGI was used for processing and preparing HTTP requests and responses while open source application programs HTML and CSS were also employed to design the front end. GBROWSE was embedded in the front end to facilitate viewing of the mapping profiles for TSS-seq reads on the genome.

**Results:** GmPromDB constitutes approximately 3700 promoter sequences with experimentally verified TSS. These promoters encompass the (-1000/+100) regions surrounding the TSS. The repository is presented to users via a web interface with the URL <http://gmpromdb.sanbi.ac.za> where the users can download their promoter(s)

of interest from a given gene or scaffold. All codes written and implemented in this project are available from the Downloads page.

**Conclusions:** Promoter sequences in GmPromDB will be instrumental to the *Glossina* research community particularly in studies focused on transcriptional control elements. GmPromDB provides a starting point for a comprehensive collection of promoters to improve the annotation of the recently sequenced *Glossina morsitans* genome and subsequent collections of newly sequenced *Glossina* species. These TSS' are being integrated with the vectorBase genome browser for *Glossina morsitans* ([www.vectorbase.org](http://www.vectorbase.org)).



## 4.1 Biological databases: a preamble

The amount of data ensuing from a genome sequencing and annotation project is enormous. The challenge thereafter lies in the ability to organize this data into a form that can be meaningful for the research community. Databases fulfill this function by managing and enabling accessibility of these data.

A database can be simply defined as an organized collection of data. In principal, all the data contained in databases must pass quality checks to ensure its integrity. Integrity checks are defined during the database and specified in a schema that captures specification of the tables and columns. To ensure non-redundancy, specific constraints are introduced during database design so that subsequent data modifications are always accurate [414].

Databases are commonly implemented using a tiered client-server architecture where the system is divided into three layers namely; data layer, logic layer and presentation layer (Figure 4.1). Whilst the logic and data layers handle majority of the data processing, the presentation layer handles display and layout processing.

The data layer constitutes the database layer and the database connection layer. The database layer processes user requests which will normally be in the form of actions such as queries or updates as well as insertions and deletions. The database connection layer connects user request to the actual database.

The logic layer is generally referred to as “the wits of the application” as it is responsible for conformance and error scrutiny. Some of these checks would be to ensure that the parameters of any request are correct followed by translating the user request to a database readable format, that is, Structured Query Language (SQL).

Finally, the presentation layer handles the output of a request that is usually format-

ted in Hypertext Markup Language (HTML) providing the end-user with the visual means of accessing and querying the system.

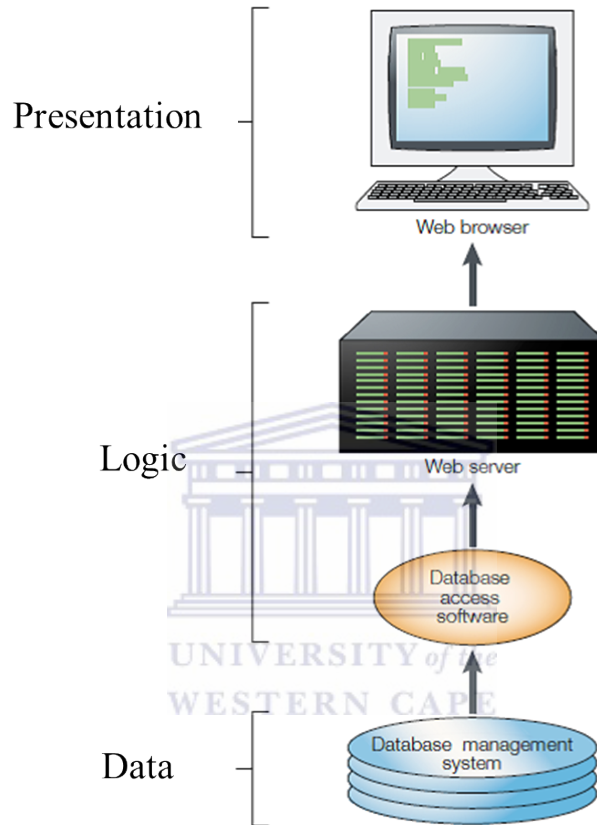
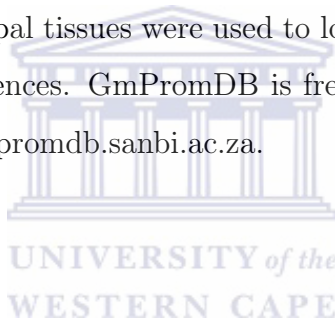


Figure 4.1: The conventional three-tier database architecture.

The Database Management System (DBMS) manages the raw data from the genome sequencing while the webserver transforms queries mediated by the database access software into hypertext mark-up language. Collectively, these are referred to as the ‘back-end’ of the system. The web browser transmits requests for data to the database and renders the responses as web pages (‘front-end’) [415].

## 4.2 Need for GmPromDB

Data accrued from both the experimental and computational large-scale analyses of transcriptional control in various organisms have been compiled as databases for example: The eukaryotic promoter database (EPDnew) [243], Database of transcription start sites (DBTSS) [416] and Mammalian promoter database (MPromDb) [417]. Information contained in these databases has been instrumental in facilitating transcription regulatory studies in corresponding organisms. GmPromDB (**G**lossina **m**orsitans **P**romoter **D**atabase) was developed as a resource to facilitate transcription regulatory studies in the newly sequenced *G.morsitans* genome. TSS positions accrued from experimental data encompassing approximately six million TSS-seq reads obtained from the larval and pupal tissues were used to locate TSS and thereby extract corresponding promoter sequences. GmPromDB is freely available to academic and non-profit users at <http://gmpromdb.sanbi.ac.za>.



## 4.3 Data assembly

Genome-wide promoter extraction was realized by an in-house pipeline which utilised high throughput NGS reads sampled from larval and pupal tissues to locate TSSs within the *G.morsitans* genome. This methodology is discussed comprehensively in chapter 2 sections 2.4 and 2.5. Tag clusters containing at least 10 uniquely mapped TSS-seq tags were considered as transcriptionally active regions. The proximal promoters +100/-1000 regions surrounding the TSS for transcriptionally active regions were extracted. They constitute 3735 promoter entries.

## 4.4 Database design

GmPromDB is a web-based resource designed using MySQL and resides on an Apache HTTP webserver. The Apache web server acts as the medium for routing requests

from the client (front end) to the database MYSQL server (back end). Open source application programs HTML and CSS were also employed to build the front end. The database contents are stored in MYSQL relational database while PERL CGI was used for processing and preparing HTTP requests and responses.

The schema embodies the three-tier architecture as described above and is summarised by the following simplified diagram;

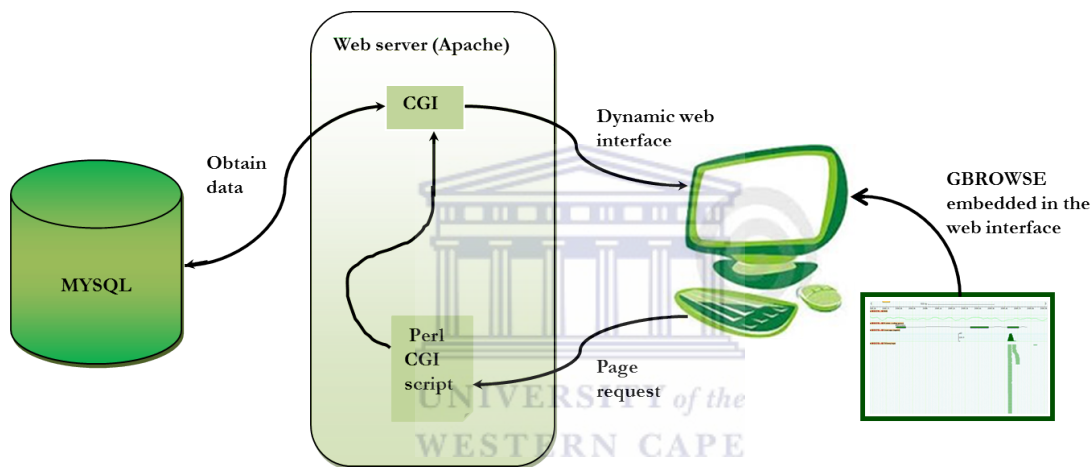


Figure 4.2: Simplified representation of the database design. The diagram illustrates the database components namely; MYSQL tables, the web server software and the dynamic user interface.

Briefly, the apache web server service runs in the background presenting the user with a query interface. After a user has input a request (page request) the web server sends the page request using standard input variables via the Common Gateway Interface (CGI). The CGI is a usual method for web server software to facilitate the generation of web content to executable files. The CGI transmits SQL requests to the database that returns the appropriate information processed and communicates to the web server using the canonical output method for display on the dynamic web interface. The mapping profile of TSS-seq reads is displayed via GBROWSE [239] embedded in the dynamic web interface.

## 4.5 Database utility

### 4.5.1 Home page

The home page provides the user with a simple introduction to *Glossina morsitans* and a brief about the database.



Figure 4.3: A snapshot of GmPromDB's home page.

### 4.5.2 Search page

The user is able to search for a promoter of a given gene of using the VectorBase [35] gene ID which begins with the acronym GMOY ID. Upon searching, the user is presented with an entry containing a summary of information pertaining to the candidate gene.

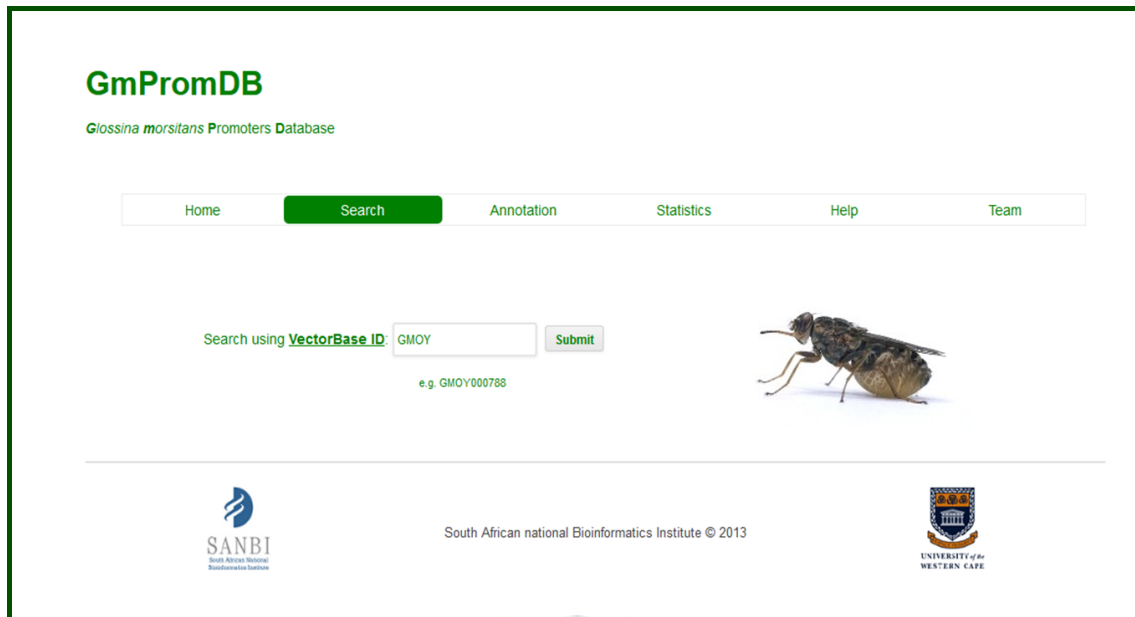


Figure 4.4: A snapshot of the search page for GmPromDB.

### 4.5.3 Search entry record

The search entry record is presented in two portions; (i) a summary of information about the gene in question and (ii) a genome browser, showing comprehensive information pertaining to the gene in question.





## Search entry record continued

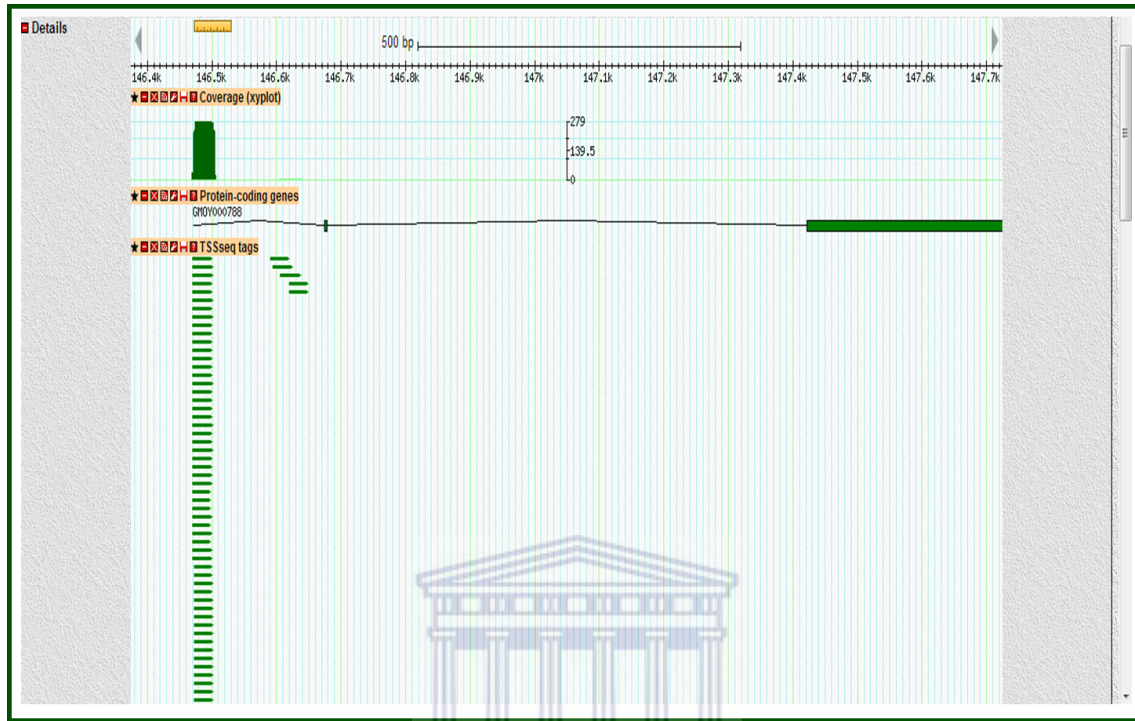


Figure 4.6: A snapshot of the search entry record. This is the second segment providing detailed information pertaining to the gene in question using GBrowse [239] embedded in the dynamic web interface.

The green box on the far right of the coverage track appears as a peak indicating a transcriptionally active region. The region denotes the TSS-seq read coverage on the genome. The corresponding TSS-seq reads appear as a vertical pile of reads right beneath the 5'UTR (black line) of the protein coding gene GMOY000788.

## 4.5.4 Downloading promoters

The user may decide to download the promoter sequence by clicking on the ‘click to download link’ embedded with the promoter sequence. The promoter sequence is displayed in a new window. Alternatively, the user may use the “downloads” page to obtain promoter sequences on a gene by gene basis or all promoters in a given scaffold (Figures 4.7 and 4.8).

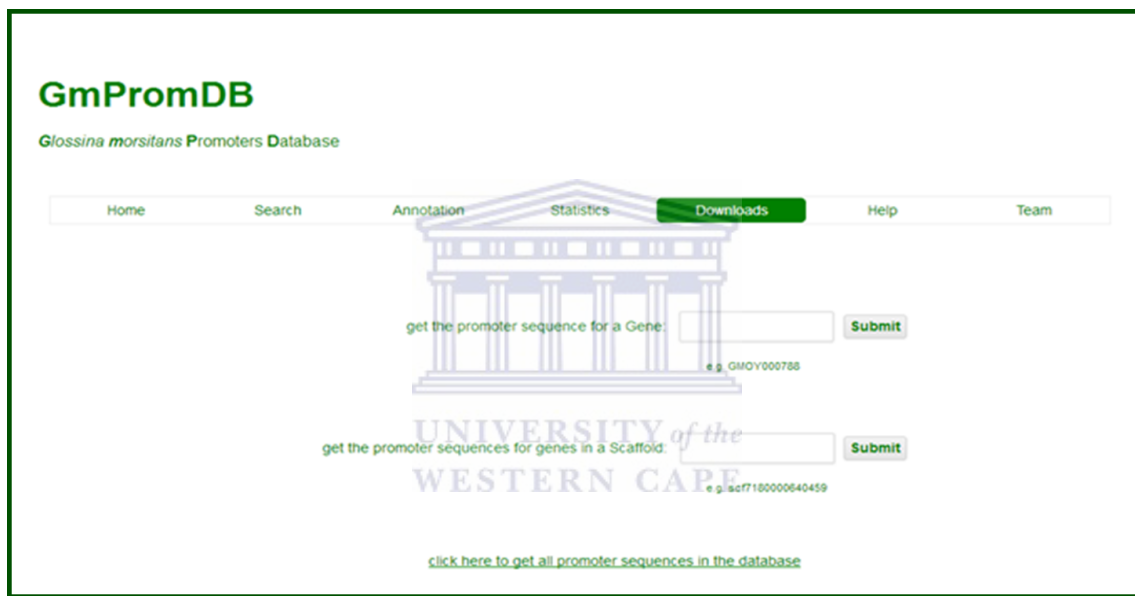


Figure 4.7: A snapshot of downloads section.

Using this interface, the user can provide the gene/scaffold name in question. Alternatively, the hyperlink at the bottom of the page allows for the download of all promoters.

Home Search Annotation Statistics Downloads Help Team

**SCF7180000640459**

GMOY000784 No TSS annotated for GMOY000784

GMOY000785 [GMOY000785\\_promoter.fa](#)

GMOY000786 No TSS annotated for GMOY000786

GMOY000787 No TSS annotated for GMOY000787

GMOY000788 [GMOY000788\\_promoter.fa](#)

GMOY000789 No TSS annotated for GMOY000789

GMOY000790 No TSS annotated for GMOY000790

\*[SCF7180000640459](#) contains 7 gene/s.

\*[SCF7180000640459](#) contains 2 gene/s with promoter sequence/s

[click here to get all promoter sequences in the scaffold](#)

Figure 4.8: A scaffold entry depicting all the genes in the scaffold. A notification of genes without experimentally determined TSS' is displayed.

By clicking on the fasta entry (for example, GMOY000788.fa), a new window displays the promoter in question. The user is then able to download the promoter. The user can download all promoters in a given scaffold.

## 4.6 Conclusion, limitations and future work

Using TSS-seq reads from two developmental tissues of the newly sequenced *G. morsitans* genome, TSSs of approximately 3700 genes were located and their corresponding promoter sequences extracted. A repository for these promoters was compiled for use by the *Glossina* research community particularly in studies focused on transcriptional control elements. Albeit modest, this resource provides a starting point for a subsequent comprehensive collection of promoters to improve the annotation of the recently sequenced *G. morsitans* genome and subsequent collection of newly sequenced *Glossina* species.

GmPromDB can provide data in future analyses to assist detailed functional studies on the *Glossina* genomes. By using GmPromDB, the core promoter structure, the presence and/or absence thereof of regulatory elements and the distribution of TSS clusters can be identified. These include studies such as those requiring promoter sequences to identify regulatory components such as TFBSs of importance for particular regulatory networks. Characterisation of regulatory networks has become an important part of genomic research in the post-genome era and promoter databases are a requisite for this sort of analysis.

There were several constraints within which we were working, for instance the state of the genome assembly which exhibits low per scaffold gene ratio. Though tag clusters mapping onto gene-less could be real TSS' but without any gene structure, some TSS' representing genes may have not been accounted for because of inconsistency in gene distribution. In addition, the manual curation effort is still on-going and as such, few genes have been assigned gene names. In future, more information furnished for the available *Glossina* gene set will facilitate incorporation of additional search criteria to enable flexibility. Secondly, since sampling was done for two developmental stages, only promoters for a subset of genes could be obtained, presumably genes preferentially expressed during development. Additional sampling of different tissues and

under different conditions will provide more promoter datasets. This will ultimately facilitate deeper analysis of the transcriptional organisation in *G.morsitans*.

The database will be updated periodically to include additional promoter data, for example, additional *Glossina* genomes as they are availed by the sequencing consortium. In addition, plans are underway to include promoter interaction networks for each of the promoters selected based on the TFBS annotation in future.



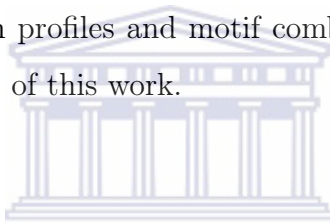
# Chapter 5

## Summary and Perspective



## 5.1 Major contributions of this work

Besides advancing the current understanding of basal transcription in the newly sequenced *G.morsitans* genome, this study has provided a foundation for transcriptional studies for the planned sequencing of *Glossina* genomes. Importantly, the study has shown that *G.morsitans* transcription initiation program exhibits emerging characteristics of metazoan core promoters where transcription initiation for most genes occurs at multiple positions in the core promoter. Initially, the canonical model of transcription initiation was believed to proceed via one TSS directing transcription using combinatorial interaction of multiple TFs. Core promoters are now classified with respect to the number of TSS and each core promoter category is associated with distinct tissue expression profiles and motif combinations. The following is an outline of major contributions of this work.



### 5.1.1 Location of TSS in the *G.morsitans* genome

A workflow that was utilized to locate TSSs on a genome-wide scale using 5' capped high throughput short reads was presented in chapter two. Given that the data availed for this study was obtained from two tissue libraries, only one-tenth of the total gene count had strong transcriptional signals and was used for subsequent analysis. Profiles obtained by mapping the short reads onto the genome were used to group core promoters into different classes, based on the number of TSS. The mapping profiles revealed that the *G.morsitans* transcriptional control program is characterized by narrow and broad core promoters similar to other metazoans. While narrow core promoters initiate transcription over several nucleotides, broad core promoters initiate over a larger genomic window. Broad core promoters were also found to exhibit preference for one initiation site (in which they were referred to as peaked) or initiate over several nucleotides with no preference for any one particular site (in which they were referred to as without peak). The trends depicted by core promoter' profiles suggests that as is the case with metazoans, majority of *G.morsitans* core promoters



are of the broad type. Accordingly, transcription initiation in *G.morsitans* genes can no longer be considered in light of the traditional model where transcription initiates at only one site.

### 5.1.2 Elucidation of *G.morsitans* core promoter architecture

Unlike mammalian core promoters, core promoters in the *G.morsitans* genome were found to exhibit propensity for the AT dinucleotides, akin to what has been observed in *D.melanogaster*. The variation in dinucleotide composition suggests a fundamental difference in global promoter architecture between mammals and insects. Unfortunately, due to the absence of high throughput NGS reads such as TSS-seq or CAGE tags for mosquitoes, we could not replicate the same analysis. Core promoter motifs in various promoter classes were shown to be present in distinct combinations showing that the core promoter motifs and their corresponding transcription factors differed across various initiation patterns. Approximately 23% of *G.morsitans* core promoters harbored a TATA-box while 26% lacked known core promoter motifs. Lack of known core promoter motifs reinforces the hypothesis that additional core promoter motifs are yet to be discovered, but the current ones are sufficient to explain the RNA pol II mediated basal transcription initiation program for majority of genes. The observation of several promoters being entangled with repeat sequences underscores the growing appreciation for the role of repeat elements in gene regulation. Information regarding core promoter architecture could be useful in designing computational models for TSS prediction in *G.morsitans* and the projected sequencing of additional *Glossina* species.

### 5.1.3 Promoter content of *G.morsitans* immunity genes

Clustering of proteins across insect proteomes allowed us to identify *G.morsitans* immunity genes. Essentially, majority of *G.morsitans* immunity gene families were

found to be systematically fewer than other insect vectors. The immune response has been presumed to be the main contributor to refractoriness, that is, innate ability of *G.morsitans* to prevent transmission of trypanosomes. As such, *G.morsitans* may employ other mechanisms to fabricate such an efficient immune response. Most of the immunity genes with transcriptional signals encode genes that function in pathways shared by developmental and immunity programs such as the toll and apoptotic pathways. In effect, majority of the corresponding over-represented transcription factors were found to have experimental evidence implicating them as regulators of immunity and development processes. The study generated *in silico* inferred hypotheses that can be tested experimentally in future. In addition, such data form a foundation for the development of gene networks to model interactions that would help unravel biological pathways that are crucial for Trypanosome transmission.



#### 5.1.4 A repository for *G.morsitans* promoters

Proximal promoter regions (-1000/+100) of genes with sufficient transcriptional signals were extracted and constitute 3735 promoters. This data was compiled into a resource named GmPromDB using a MYSQL relational database. Databases are useful resources for mining and exploration of data, for example data obtained from genome sequencing projects. The creation of GmPromDB was the first step towards a systematic analysis of *G.morsitans* and promoters of *Glossina* genomes that are yet to be sequenced. GmPromDB is also a useful resource for the insect research community specifically studies targeted at comparative studies of transcriptional regulatory elements. The experimentally verified TSS locations derived from this study are currently being integrated into *G.morsitans* genome browser at VectorBase ([www.vectorbase.org](http://www.vectorbase.org)).

## 5.2 Perspective

In the recent years, emerging models of transcription regulation have changed the conventional understanding of transcription initiation programs generating more interest on transcription regulation studies. The work presented herein can be viewed as a foundation for empowering *Glossina* researchers to reach a better understanding of the fundamental complex biological processes involved in *G.morsitans* transcription initiation. This study has provided a basis for future exploration of transcriptional control mechanisms in *G.morsitans* and the projected sequencing of additional *Glossina* species. Some future plans are outlined below.

### 5.2.1 Improvement of the *G.morsitans* genome assembly and annotation

The genome sequence and set of 12,220 predicted genes will be refined over time as annotations are improved and genes are functionally characterized. Complete annotations of genes in terms of assigning gene names and/or descriptions and curation of gene models will expedite future targeted analysis. The diversity of predicted genes and gene products will serve as the basis for additional experimental work. This will help unravel molecular mechanisms underlying important aspects of Tsetse biology, specifically the transmission cycle. Moreover, a well annotated *G.morsitans* genome sequence will form a benchmark for comparative studies with additional *Glossina* genomes. Further assembly and/or physical mapping will enable detailed analysis of transcriptional studies. For instance, the fragmented status of the genome and skewed scaffold to gene ratio complicated the analysis. This is because TSS-seq clusters mapping on the periphery of scaffolds or scaffolds without an annotated gene could not be associated with protein-coding genes. A further assembled genome would facilitate analysis of expression clusters. In eukaryotes, gene order in eukaryotic genomes is not random and genes with similar expression profiles tend to cluster as local expression clusters. With regard to transcriptional studies, analysis of expression clusters

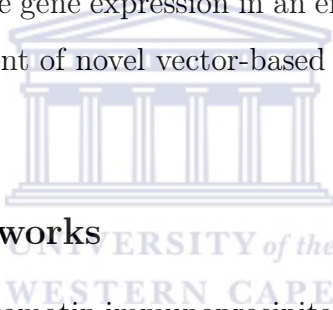
may shed light on promoter usage and provide answers to questions such as whether co-expressed genes in *G.morsitans* are organized as clusters and if so do they share promoters or alternative promoters. For example, in this study some genes had several TSS-seq tag clusters some which may have representatives of alternative promoters. Further assembly and/or physical mapping would also allow assessment of contribution of the cis-regulatory elements that may be present in the intergenic regions.

### 5.2.2 Inclusion of additional datasets

Deeper sampling of different tissues under different conditions and/or at different time points using integrated approaches such as RNA-seq, TSS-seq and ChIP-seq would enable integrative interpretation of transcriptome data. This includes analysis of differential promoter usage in different tissues and at specific conditions as well as epigenetic architecture of transcription initiation. Transcriptional studies in *D.melanogaster* have shown that distinct initiation patterns and motif usage are associated with different tissues and/or time points. Besides providing information on epigenetic patterns and differential promoter usage, integrative analysis will provide a wealth of information about putative alternative promoters which were not explored in the current study. Alternative promoters may have insightful downstream effects such as the diversification of a gene's isoforms, an increase in the complexity of a gene's architecture and possibly, an expansion of the biochemical role of a gene's function. Identification of the variety of mRNAs generated by alternative promoters will increase the protein repertoire. Integrative analysis with an improved genome assembly may also provide data on non-coding RNA. Specifically, short reads mapping on intergenic sites may represent TSS for non-coding RNAs that have important roles in gene regulation. For example, in this study approximately one third (1101/3134) TSS-seq tag clusters were identified in intergenic regions and may be candidates for non-coding RNAs.

### 5.2.3 Experimental validation of predicted TFBSs

Experimental confirmation is still the best form of TFBSs validation because provides accurate information about the inferred biological function of a TFBS together the identity of its corresponding TF. Biochemical assays such as gel shift mobility and DNase footprinting assays on selected computationally predicted targets would help ascertain the presence and/or absence of of the core promoter elements. Large-scale assays such as ChIP-sequencing may expedite such analysis for example, by being able to determine the binding sites of the TATA-binding protein on a genome-wide scale. Experimental validation of promoter profiles of immunity genes for instance may be facilitated by analysis of promoter reporter assays. These promoters may be used to drive anti-trypanosome gene expression in an efficient time and tissue-specific mode aiding in the development of novel vector-based control strategies.



### 5.2.4 Regulatory networks

Biochemical assays such as chromatin immunoprecipitation coupled with gene expression profiling, and computational methods will enable construction of blueprints for the initiation and maintenance of complex cellular processes. Of specific interest are processes that are involved during Tsetse-Trypanosome interactions such as immunity (required for Trypanosome elimination) and salivary gland processes (required for successful transmission). By determining promoter occupancy of all promoter regions of promoters of genes preferentially expressed during the aforementioned processes, TFs obtained may offer some insight into the global regulatory network of these processes. In addition such analysis may also unravel the role of cis-acting elements in these promoters. Improved predictions of such networks may find widespread application towards efforts to delineate the impact of Trypanosome establishment on cellular responses in the Tsetse flies. This information may be exploited further to design novel methods for halting the transmission cycle.

## 5.3 Final remarks

The limitations notwithstanding, in this study a foundation has been laid for future detailed functional studies. In essence, a TSS classification algorithm using TSS-seq data has been developed and five questions have been answered regarding transcriptional control in *G.morsitans*:

- 1) Does *G.morsitans* transcriptional program exhibit emerging characteristics of meta-zoan promoters?
- 2) Are the core promoters of *G.morsitans* and mammals fundamentally different in terms of nucleotide composition?
- 3) Do *G.morsitans* core promoters utilize canonical core promoter motifs?
- 4) Is there a variation in core promoter motif co-occurrence across promoter classes?
- 5) Do *G.morsitans* promoters of immunity genes harbor transcription factor binding sites with transcription factors implicated as regulators of the immune response?

The answer to all these questions is yes. However, since they are all based on *in silico* inferred hypothesis, experimental verifications in the future will expedite current understanding of *G.morsitans* transcription mechanisms.

# Bibliography

- [1] Petersen FT, Meier R, Kutty SN, Wiegmann BM: The phylogeny and evolution of host choice in the Hippoboscoidea (Diptera) as reconstructed using four molecular markers. *Mol. Phylogenet. Evol.* 2007, 45:111122.
- [2] Simarro P, Cecchi G, Paone M, Franco J, Diarra A, Ruiz J, Fevre E, Courtin F, Mattioli R, Jannin J: The Atlas of human African trypanosomiasis: a contribution to global mapping of neglected tropical diseases. *Int. J. Health Geogr.* 2010, 9:57.
- [3] Simarro PP, Diarra A, Ruiz Postigo JA, Franco JR, Jannin JG: The human African trypanosomiasis control and surveillance programme of the World Health Organization 2000-2009: the way forward. *PLoS Negl. Trop. Dis.* 2011, 5:e1007.
- [4] Hotez PJ, Kamath A: Neglected tropical diseases in sub-saharan Africa: review of their prevalence, distribution, and disease burden. *PLoS Negl. Trop. Dis.* 2009, 3:e412.
- [5] Franco J, Simarro P, Diarra, Ruiz Postigo, Jannin: Diversity of human African trypanosomiasis epidemiological settings requires fine-tuning control strategies to facilitate disease elimination. *Res. Rep. Trop. Med.* 2013:1.
- [6] Vickerman K: Developmental cycles and biology of pathogenic trypanosomes. *Br. Med. Bull.* 1985, 41:105114.

- [7] Pays E, Lips S, Nolan D, Vanhamme L, Pz-Morga D: The VSG expression sites of *Trypanosoma brucei*: multipurpose tools for the adaptation of the parasite to mammalian hosts. *Mol. Biochem. Parasitol.* 2001, 114:116.
- [8] Ferreira F, Cano J, Furtado A, Ndong-Mabale N, Ndong-Asumu P, Benito A, Pinto J, Afonso MO, Seixas J, Atougua J, Centeno-Lima S: An alternative approach to detect *Trypanosoma* in *Glossina* (Diptera, Glossinidae) without dissection. *J. Infect. Dev. Ctries.* 2008, 2:6367.
- [9] Stuart K, Brun R, Croft S, Fairlamb A, Grtler RE, McKerrow J, Reed S, Tarleton R: Kinetoplastids: related protozoan pathogens, different diseases. *J. Clin. Invest.* 2008, 118:13011310.
- [10] Roditi I, Lehane MJ: Interactions between trypanosomes and tsetse flies. *Curr. Opin. Microbiol.* 2008, 11:345351.
- [11] Abubakar LU, Bulimo WD, Mula FJ, Osir EO: Molecular characterization of a tsetse fly midgut proteolytic lectin that mediates differentiation of African trypanosomes. *Insect Biochem. Mol. Biol.* 2006, 36:344352.
- [12] Hao Z, Kasumba I, Lehane MJ, Gibson WC, Kwon J, Aksoy S: Tsetse immune responses and trypanosome transmission: implications for the development of tsetse-based strategies to reduce trypanosomiasis. *Proc. Natl. Acad. Sci. U. S. A.* 2001, 98:1264812653.
- [13] Boulanger N, Brun R, Ehret-Sabatier L, Kunz C, Bulet P: Immunopeptides in the defense reactions of *Glossina morsitans* to bacterial and *Trypanosoma brucei brucei* infections. *Insect Biochem. Mol. Biol.* 2002, 32:369375.
- [14] Lehane MJ, Aksoy S, Gibson W, Kerhornou A, Berriman M, Hamilton J, Soares MB, Bonaldo MF, Lehane S, Hall N: Adult midgut expressed sequence tags from the tsetse fly *Glossina morsitans morsitans* and expression analysis of putative immune response genes. *Genome Biol.* 2003, 4:R63.



- [15] Hu C, Aksoy S: Innate immune responses regulate trypanosome parasite infection of the tsetse fly *Glossina morsitans morsitans*. *Mol. Microbiol.* 2006, 60:11941204.
- [16] Milligan PJ, Maudlin I, Welburn SC: Trypanozoon: infectivity to humans is linked to reduced transmissibility in tsetse. II. Genetic mechanisms. *Exp. Parasitol.* 1995, 81:409415.
- [17] Welburn SC, Maudlin I, Milligan PJ: Trypanozoon: infectivity to humans is linked to reduced transmissibility in tsetse. I. Comparison of human serum-resistant and human serum-sensitive field isolates. *Exp. Parasitol.* 1995, 81:404408.
- [18] MacLeod ET, Maudlin I, Darby AC, Welburn SC: Antioxidants promote establishment of trypanosome infections in tsetse. *Parasitology* 2007, 134:827831.
- [19] Hendry KA, Vickerman K: The requirement for epimastigote attachment during division and metacyclogenesis in *Trypanosoma congolense*. *Parasitol. Res.* 1988, 74:403408.
- [20] Van Den Abbeele J, Caljon G, De Ridder K, De Baetselier P, Coosemans M: *Trypanosoma brucei* modifies the tsetse salivary composition, altering the fly feeding behavior that favors parasite transmission. *PLoS Pathog.* 2010, 6:e1000926.
- [21] Steverding D: The development of drugs for treatment of sleeping sickness: a historical review. *Parasit. Vectors* 2010, 3:15.
- [22] Barrett MP, Boykin DW, Brun R, Tidwell RR: Human African trypanosomiasis: pharmacological re-engagement with a neglected disease. *Br. J. Pharmacol.* 2007, 152:11551171.
- [23] Barry JD, McCulloch R: Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite. *Adv. Parasitol.* 2001, 49:170.

- [24] Infectious Diseases of Humans: Dynamics and Control - Roy M. Anderson, Robert M. May - Google Books [<http://books.google.co.za/books?id=HT0-xXBguQCpg=PA39lpq=PA39dq>].
- [25] Berriman M, Ghedin E, Hertz-Fowler C, et al.: The genome of the African trypanosome *Trypanosoma brucei*. *Science* 2005, 309:416422.
- [26] Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, Quail MA, Chukualim B, Capewell P, MacLeod A, Melville SE, Gibson W, Barry JD, Berriman M, Hertz-Fowler C: The genome sequence of *Trypanosoma brucei gambiense*, causative agent of chronic human african trypanosomiasis. *PLoS Negl. Trop. Dis.* 2010, 4:e658.
- [27] Amaro RE, Schnauffer A, Interthal H, Hol W, Stuart KD, McCammon JA: Discovery of drug-like inhibitors of an essential RNA-editing ligase in *Trypanosoma brucei*. *Proc. Natl. Acad. Sci.* 2008, 105:1727817283.
- [28] A proposal for tsetse fly (*Glossina*) genome projects [<https://www.vectorbase.org/projects/proposal-tsetse-fly-glossina-genome-projects>].
- [29] Aksoy S, Maudlin I, Dale C, Robinson AS, O'Neill SL: Prospects for control of African trypanosomiasis by tsetse vector manipulation. *Trends Parasitol.* 2001, 17:2935.
- [30] Geiger A, Ravel S, Frutos R, Cuny G: *Sodalis glossinidius* (Enterobacteriaceae) and vectorial competence of *Glossina palpalis gambiense* and *Glossina morsitans morsitans* for *Trypanosoma congolense savannah* type. *Curr. Microbiol.* 2005, 51:3540.
- [31] Cheng Q, Aksoy S: Tissue tropism, transmission and expression of foreign genes in vivo in midgut symbionts of tsetse flies. *Insect Mol. Biol.* 1999, 8:125132.
- [32] Haines LR, Hancock REW, Pearson TW: Cationic antimicrobial peptide killing of African trypanosomes and *Sodalis glossinidius*, a bacterial symbiont of the

insect vector of sleeping sickness. *Vector Borne Zoonotic Dis. Larchmt. N* 2003, 3:175186.

[33] De Vooght L, Caljon G, Stijlemans B, De Baetselier P, Coosemans M, Van den Abbeele J: Expression and extracellular release of a functional anti-trypanosome Nanobody in *Sodalis glossinidius*, a bacterial symbiont of the tsetse fly. *Microb. Cell Factories* 2012, 11:23.

[34] Venter JC, Adams MD, Myers EW, et al.: The sequence of the human genome. *Science* 2001, 291:13041351.

[35] *Glossina morsitans* | VectorBase [<https://www.vectorbase.org/organisms/glossina-morsitans>].

[36] Adams MD, Celniker SE, Holt RA, et al.: The genome sequence of *Drosophila melanogaster*. *Science* 2000, 287:21852195.

[37] Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M: Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques* 2008, 45:8194.

[38] RepeatMasker Home Page [<http://www.repeatmasker.org/>].

[39] Curwen V, Eyraas E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M: The Ensembl automatic gene annotation system. *Genome Res.* 2004, 14:942950.

[40] CAPvectorBase-Cap @ EnsemblGenomes [<http://vectorbase-cap.ensemblgenomes.org/?q=cap>].

[41] Blaxter M: Genetics. Revealing the dark matter of the genome. *Science* 2010, 330:17581759.

[42] Juven-Gershon T, Hsu J-Y, Theisen JW, Kadonaga JT: The RNA polymerase II core promoter the gateway to transcription. *Curr. Opin. Cell Biol.* 2008, 20:253259.

- [43] Wray GA: The Evolution of Transcriptional Regulation in Eukaryotes. *Mol. Biol. Evol.* 2003, 20:13771419.
- [44] Juven-Gershon T, Hsu J-Y, Kadonaga JT: Perspectives on the RNA polymerase II core promoter. *Biochem. Soc. Trans.* 2006, 34:10471050.
- [45] Narlikar GJ, Fan H-Y, Kingston RE: Cooperation between complexes that regulate chromatin structure and transcription. *Cell* 2002, 108:475487.
- [46] Lemon B, Tjian R: Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* 2000, 14:25512569.
- [47] Lee TI, Young RA: Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.* 2000, 34:77137.
- [48] Maston GA, Evans SK, Green MR: Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* 2006, 7:2959.
- [49] Blackwood EM, Kadonaga JT: Going the distance: a current view of enhancer action. *Science* 1998, 281:6063.
- [50] Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E: A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* 2003, 12:17251735.
- [51] Privalsky ML: The role of corepressors in transcriptional regulation by nuclear hormone receptors. *Annu. Rev. Physiol.* 2004, 66:315360.
- [52] Ogbourne S, Antalis TM: Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem. J.* 1998, 331 ( Pt 1):114.
- [53] Lenhard B, Sandelin A, Carninci P: Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* 2012, 13:233245.

- [54] Dikstein R: The unexpected traits associated with core promoter elements. *Transcription* 2011, 2:201206.
- [55] Sawadogo M, Roeder RG: Interaction of a gene-specific transcription factor with the adenovirus major late promoter upstream of the TATA box region. *Cell* 1985, 43:165175.
- [56] Nakatani Y, Horikoshi M, Brenner M, Yamamoto T, Besnard F, Roeder RG, Freese E: A downstream initiation element required for efficient TATA box binding and in vitro function of TFIID. *Nature* 1990, 348:8688.
- [57] Zhou QY, Li C, Civelli O: Characterization of gene organization and promoter region of the rat dopamine D1 receptor gene. *J. Neurochem.* 1992, 59:18751883.
- [58] Wang JC, Van Dyke MW: Sp1, USF, and GAL4 activate transcription independently of TFIID-downstream promoter interactions. *Biochim. Biophys. Acta* 1994, 1218:308314.
- [59] Kaufmann J, Smale ST: Direct recognition of initiator elements by a component of the transcription factor IID complex. *Genes Dev.* 1994, 8:821829.
- [60] Purnell BA, Emanuel PA, Gilmour DS: TFIID sequence recognition of the initiator and sequences farther downstream in *Drosophila* class II genes. *Genes Dev.* 1994, 8:830842.
- [61] Oelgeschlager T, Chiang CM, Roeder RG: Topology and reorganization of a human TFIID-promoter complex. *Nature* 1996, 382:735738.
- [62] Verrijzer CP, Tjian R: TAFs mediate transcriptional activation and promoter selectivity. *Trends Biochem. Sci.* 1996, 21:338342.
- [63] Deng W, Roberts SGE: A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev.* 2005, 19:24182423.
- [64] Gazit K, Moshonov S, Elfakess R, Sharon M, Mengus G, Davidson I, Dikstein R: TAF4/4b x TAF12 displays a unique mode of DNA binding and is

required for core promoter function of a subset of genes. *J. Biol. Chem.* 2009, 284:2628626296.

- [65] Smale ST: Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim. Biophys. Acta* 1997, 1351:7388.
- [66] Jin VX, Singer GAC, Agosto-Pz FJ, Liyanarachchi S, Davuluri RV: Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. *BMC Bioinformatics* 2006, 7:114.
- [67] Tokusumi Y, Ma Y, Song X, Jacobson RH, Takada S: The new core promoter element XCPE1 (X Core Promoter Element 1) directs activator-, mediator-, and TATA-binding protein-dependent but TFIID-independent RNA polymerase II transcription from TATA-less promoters. *Mol. Cell. Biol.* 2007, 27:18441858.
- [68] Deng W, Roberts SGE: TFIIB and the regulation of transcription by RNA polymerase II. *Chromosoma* 2007, 116:417429.
- [69] Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH: New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev.* 1998, 12:3444.
- [70] FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C: Comparative genomics of *Drosophila* and human core promoters. *Genome Biol.* 2006, 7:R53.
- [71] Butler JEF, Kadonaga JT: The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* 2002, 16:25832592.
- [72] Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engstrm PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai

- J, Bajic VB, Hume DA, Hayashizaki Y: Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 2006, 38:626635.
- [73] Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A: A code for transcription initiation in mammalian genomes. *Genome Res.* 2007, 18:112.
- [74] Lim CY, Santoso B, Boulay T, Dong E, Ohler U, Kadonaga JT: The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev.* 2004, 18:16061617.
- [75] Burke TW, Kadonaga JT: *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.* 1996, 10:711724.
- [76] Burke TW, Kadonaga JT: The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII60 of *Drosophila*. *Genes Dev.* 1997, 11:30203031.
- [77] Kutach AK, Kadonaga JT: The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol. Cell. Biol.* 2000, 20:47544764.
- [78] Smale ST, Kadonaga JT: The RNA polymerase II core promoter. *Annu. Rev. Biochem.* 2003, 72:449479.
- [79] Juven-Gershon T, Kadonaga JT: Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.* 2010, 339:225229.
- [80] Rach EA, Yuan H-Y, Majoros WH, Tomancak P, Ohler U: Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the *Drosophila* genome. *Genome Biol.* 2009, 10:R73.
- [81] Hawley DK, McClure WR: Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Res.* 1983, 11:22372255.

- [82] Ross W, Gosink KK, Salomon J, Igarashi K, Zou C, Ishihama A, Severinov K, Gourse RL: A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science* 1993, 262:14071413.
- [83] Dombroski AJ, Walter WA, Record MT Jr, Siegele DA, Gross CA: Polypeptides containing highly conserved regions of transcription initiation factor sigma 70 exhibit specificity of binding to promoter DNA. *Cell* 1992, 70:501512.
- [84] Blatter EE, Ross W, Tang H, Gourse RL, Ebright RH: Domain organization of RNA polymerase alpha subunit: C-terminal 85 amino acids constitute a domain capable of dimerization and DNA binding. *Cell* 1994, 78:889896.
- [85] Jacob F, Perrin D, Sanchez C, Monod J: (Operon: a group of genes with the expression coordinated by an operator). *Comptes Rendus Hebd. Sces Acade Sci.* 1960, 250:17271729.
- [86] Blumenthal T: Operons in eukaryotes. *Brief. Funct. Genomic. Proteomic.* 2004, 3:199211.
- [87] Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T: Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* 2007, 8:67.
- [88] Yamamoto YY, Ichida H, Abe T, Suzuki Y, Sugano S, Obokata J: Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucleic Acids Res.* 2007, 35:62196226.
- [89] Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA: Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.* 2007, 8:424436.
- [90] Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J: Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. *Plant J. Cell Mol. Biol.* 2009, 60:350362.



- [91] Deaton AM, Bird A: CpG islands and the regulation of transcription. *Genes Dev.* 2011, 25:10101022.
- [92] Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, Yang L, Boley N, Andrews J, Kaufman TC, Graveley BR, Bickel PJ, Carninci P, Carlson JW, Celniker SE: Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Res.* 2010, 21:182192.
- [93] Han L, Zhao Z: Comparative analysis of CpG islands in four fish genomes. *Comp. Funct. Genomics* 2008:565631.
- [94] Sharif J, Endo TA, Toyoda T, Koseki H: Divergence of CpG island promoters: a consequence or cause of evolution? *Dev. Growth Differ.* 2010, 52:545554.
- [95] Okamura K, Yamashita R, Takimoto N, Nishitsuji K, Suzuki Y, Kusakabe TG, Nakai K: Profiling ascidian promoters as the primordial type of vertebrate promoter. *BMC Genomics* 2011, 12 Suppl 3:S7.
- [96] King MC, Wilson AC: Evolution at two levels in humans and chimpanzees. *Science* 1975, 188:107116.
- [97] Rifkin SA, Kim J, White KP: Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat. Genet.* 2003, 33:138144.
- [98] Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP: Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* 2006, 440:242245.
- [99] Khaitovich P, Enard W, Lachmann M, Po S: Evolution of primate gene expression. *Nat. Rev. Genet.* 2006, 7:693702.
- [100] Kliebenstein DJ, West MAL, van Leeuwen H, Kim K, Doerge RW, Michelmore RW, St Clair DA: Genomic survey of gene expression diversity in *Arabidopsis thaliana*. *Genetics* 2006, 172:11791189.

- [101] Landry CR, Oh J, Hartl DL, Cavalieri D: Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* 2006, 366:343351.
- [102] Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M: Divergence of transcription factor binding sites across related yeast species. *Science* 2007, 317:815819.
- [103] Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M: Genetic analysis of variation in transcription factor binding in yeast. *Nature* 2010, 464:11871191.
- [104] Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, Hong M-Y, Karczewski KJ, Huber W, Weissman SM, Gerstein MB, Korb J, Snyder M: Variation in transcription factor binding among humans. *Science* 2010, 328:232235.
- [105] Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT: Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science* 2010, 328:10361040.
- [106] Nagarajan M, Veyrieras J-B, de Dieuleveult M, Bottin H, Fehrman S, Abraham A-L, Croze S, Steinmetz LM, Gidrol X, Yvert G: Natural Single-Nucleosome Epi-Polymorphisms in Yeast. *PLoS Genet.* 2010, 6:e1000913.
- [107] Field Y, Fondufe-Mittendorf Y, Moore IK, Mieczkowski P, Kaplan N, Lubling Y, Lieb JD, Widom J, Segal E: Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nat. Genet.* 2009, 41:438445.
- [108] Tirosh I, Sigal N, Barkai N: Divergence of nucleosome positioning between two closely related yeast species: genetic basis and functional consequences. *Mol. Syst. Biol.* 2010, 6.

- [109] Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ: The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol.* 2010, 8:e1000414.
- [110] Van Heeringen SJ, Akhtar W, Jacobi UG, Akkers RC, Suzuki Y, Veenstra GJC: Nucleotide composition-linked divergence of vertebrate core promoter architecture. *Genome Res.* 2011, 21:410421.
- [111] Sinha S, Siggia ED: Sequence turnover and tandem repeats in cis-regulatory modules in drosophila. *Mol. Biol. Evol.* 2005, 22:874885.
- [112] Ludwig MZ, Patel NH, Kreitman M: Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Dev. Camb. Engl.* 1998, 125:949958.
- [113] Ludwig MZ, Bergman C, Patel NH, Kreitman M: Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 2000, 403:564567.
- [114] Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M: Functional evolution of a cis-regulatory module. *PLoS Biol.* 2005, 3:e93.
- [115] Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CAM: Heterotachy in mammalian promoter evolution. *PLoS Genet.* 2006, 2:e30.
- [116] Dermitzakis ET, Clark AG: Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 2002, 19:11141121.
- [117] Stone JR, Wray GA: Rapid evolution of cis-regulatory sequences via local point mutations. *Mol. Biol. Evol.* 2001, 18:17641770.
- [118] Huang W, Nevins JR, Ohler U: Phylogenetic simulation of promoter evolution: estimation and modeling of binding site turnover events and assessment of their impact on alignment tools. *Genome Biol.* 2007, 8:R225.

- [119] Frith MC, Ponjavic J, Fredman D, Kai C, Kawai J, Carninci P, Hayashizaki Y, Hayashizaki Y, Sandelin A: Evolutionary turnover of mammalian transcription start sites. *Genome Res.* 2006, 16:713722.
- [120] Main BJ, Smith AD, Jang H, Nuzhdin SV: Transcription start site evolution in *Drosophila*. *Mol. Biol. Evol.* 2013, 30:19661974.
- [121] Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler ELW, Zapp ML, Weng Z, Zamore PD: Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. *Science* 2008, 320:10771081.
- [122] Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, Surani MA, Sakaki Y, Sasaki H: Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 2008, 453:539543.
- [123] Conley AB, Miller WJ, Jordan IK: Human cis natural antisense transcripts initiated by transposable elements. *Trends Genet. TIG* 2008, 24:5356.
- [124] Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest ARR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P: The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* 2009, 41:563571.
- [125] Jordan IK, Rogozin IB, Glazko GV, Koonin EV: Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet. TIG* 2003, 19:6872.
- [126] Rockman MV, Wray GA: Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.* 2002, 19:19912004.
- [127] Betr, Thornton K, Long M: Retroposed new genes out of the X in *Drosophila*. *Genome Res.* 2002, 12:18541859.

- [128] Vences MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ: Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* 2009, 324:12131216.
- [129] Molina C, Grotewold E: Genome wide analysis of Arabidopsis core promoters. *BMC Genomics* 2005, 6:25.
- [130] Faulkner GJ, Carninci P: Altruistic functions for selfish DNA. *Cell Cycle Georgetown. Tex* 2009, 8:28952900.
- [131] Plessy C, Bertin N, Takahashi H, Simone R, Salimullah M, Lassmann T, Vitezic M, Severin J, Olivarius S, Lazarevic D, Hornig N, Orlando V, Bell I, Gao H, Dumais J, Kapranov P, Wang H, Davis CA, Gingeras TR, Kawai J, Daub CO, Hayashizaki Y, Gustincich S, Carninci P: Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat. Methods* 2010, 7:528534.
- [132] ENCODE Project Consortium: The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004, 306:636640.
- [133] Kurokawa R, Rosenfeld MG, Glass CK: Transcriptional regulation through non-coding RNAs and epigenetic modifications. *RNA Biol.* 2009, 6:233236.
- [134] Jacquier A: The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.* 2009, 10:833844.
- [135] Carninci P: RNA dust: where are the genes? *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* 2010, 17:5159.
- [136] Valen E, Preker P, Andersen PR, Zhao X, Chen Y, Ender C, Dueck A, Meister G, Sandelin A, Jensen TH: Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nat. Struct. Mol. Biol.* 2011, 18:10751082.

- [137] Kurokawa R: Promoter-associated long noncoding RNAs repress transcription through a RNA binding protein TLS. *Adv. Exp. Med. Biol.* 2011, 722:196208.
- [138] Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otilar RP, Myers RM: An abundance of bidirectional promoters in the human genome. *Genome Res.* 2004, 14:6266.
- [139] Lin JM, Collins PJ, Trinklein ND, Fu Y, Xi H, Myers RM, Weng Z: Transcription factor binding and modified histones in human bidirectional promoters. *Genome Res.* 2007, 17:818827.
- [140] Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Munster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM: Bidirectional promoters generate pervasive transcription in yeast. *Nature* 2009, 457:10331037.
- [141] Wang Q, Wan L, Li D, Zhu L, Qian M, Deng M: Searching for bidirectional promoters in *Arabidopsis thaliana*. *BMC Bioinformatics* 2009, 10 Suppl 1:S29.
- [142] Lecanidou R, Papantonis A: Modeling bidirectional transcription using silkworm chorion gene promoters. *Organogenesis* 2010, 6:5458.
- [143] Koyanagi KO, Hagiwara M, Itoh T, Gojobori T, Imanishi T: Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene* 2005, 353:169176.
- [144] Neil H, Malabat C, d Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A: Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* 2009, 457:10381042.
- [145] Adachi N, Lieber MR: Bidirectional gene organization: a common architectural feature of the human genome. *Cell* 2002, 109:807809.
- [146] Li Y-Y, Yu H, Guo Z-M, Guo T-Q, Tu K, Li Y-X: Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS Comput. Biol.* 2006, 2:e74.

- [147] Yang MQ, Taylor J, Elnitski L: Comparative analyses of bidirectional promoters in vertebrates. *BMC Bioinformatics* 2008, 9 Suppl 6:S9.
- [148] Xu C, Chen J, Shen B: The preservation of bidirectional promoter architecture in eukaryotes: what is the driving force? *BMC Syst. Biol.* 2012, 6 Suppl 1:S21.
- [149] Rombauts S: Computational Approaches to Identify Promoters and cis-Regulatory Elements in Plant Genomes. *PLANT Physiol.* 2003, 132:11621176.
- [150] Zhang C, Zhao X, Zhang M?: [http://rulai.cshl.edu/reprints/bioinfo4gen\\_chapter12\\_zhang.pdf](http://rulai.cshl.edu/reprints/bioinfo4gen_chapter12_zhang.pdf).
- [151] Abeel T, Saeys Y, Rouz Van de Peer Y: ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. *Bioinforma. Oxf. Engl.* 2008, 24:i2431.
- [152] Davuluri RV: Application of FirstEF to find promoters and first exons in the human genome. *Curr. Protoc. Bioinforma.* Ed. Board Andreas Baxevanis Al 2003, Chapter 4:Unit4.7.
- [153] Reese MG: Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput. Chem.* 2001, 26:5156.
- [154] Prestridge DS: Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* 1995, 249:923932.
- [155] Kondrakhin YV, Kel AE, Kolchanov NA, Romashchenko AG, Milanesi L: Eukaryotic promoter recognition by binding sites for transcription factors. *Comput. Appl. Biosci. CABIOS* 1995, 11:477488.
- [156] Hutchinson GB: The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput. Appl. Biosci. CABIOS* 1996, 12:391398.
- [157] Claverie J-M, Bougueleret L: Heuristic informational analysis of sequences. *Nucleic Acids Res.* 1986, 14:179196.
- [158] Zhang MQ: Identification of human gene core promoters in silico. *Genome Res.* 1998, 8:319326.

- [159] Scherf M, Klingenhoff A, Werner T: Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* 2000, 297:599606.
- [160] Bajic VB, Tan SL, Suzuki Y, Sugano S: Promoter prediction analysis on the whole human genome. *Nat. Biotechnol.* 2004, 22:14671473.
- [161] Suzuki Y, Sugano S: Construction of full-length-enriched cDNA libraries. The oligo-capping method. *Methods Mol. Biol.* Clifton NJ 2001, 175:143153.
- [162] Carninci P, Waki K, Shiraki T, Konno H, Shibata K, Itoh M, Aizawa K, Arakawa T, Ishii Y, Sasaki D, Bono H, Kondo S, Sugahara Y, Saito R, Osato N, Fukuda S, Sato K, Watahiki A, Hirozane-Kishikawa T, Nakamura M, Shibata Y, Yasunishi A, Kikuchi N, Yoshiki A, Kusakabe M, Gustincich S, Beisel K, Pavan W, Aidinis V, Nakagawara A, Held WA, Iwata H, Kono T, Nakauchi H, Lyons P, Wells C, Hume DA, Fagiolini M, Hensch TK, Brinkmeier M, Camper S, Hirota J, Mombaerts P, Muramatsu M, Okazaki Y, Kawai J, Hayashizaki Y: Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* 2003, 13:12731289.
- [163] Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y: Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.* 2003, 100:1577615781.
- [164] Carninci P, Hayashizaki Y: High-efficiency full-length cDNA cloning. *Methods Enzymol.* 1999, 303:1944.
- [165] Tsuchihara K, Suzuki Y, Wakaguri H, Irie T, Tanimoto K, Hashimoto S -i., Matsushima K, Mizushima-Sugano J, Yamashita R, Nakai K, Bentley D, Esumi H, Sugano S: Massive transcriptional start site analysis of human genes in hypoxia cells. *Nucleic Acids Res.* 2009, 37:22492263.



- [166] Suzuki Y, Sugano S: Construction of a full-length enriched and a 5-end enriched cDNA library using the oligo-capping method. *Methods Mol. Biol.* Clifton NJ 2003, 221:7391.
- [167] Elnitski L, Jin VX, Farnham PJ, Jones SJM: Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. *Genome Res.* 2006, 16:14551464.
- [168] Wasserman WW, Sandelin A: Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 2004, 5:276287.
- [169] Cornish-Bowden A: Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.* 1985, 13:30213030.
- [170] Stormo GD: Consensus patterns in DNA. *Methods Enzymol.* 1990, 183:211221.
- [171] King OD, Roth FP: A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.* 2003, 31:e116.
- [172] Hertz GZ, Stormo GD: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinforma. Oxf. Engl.* 1999, 15:563577.
- [173] Lawrence CE, Reilly AA: An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* 1990, 7:4151.
- [174] Hertz GZ, Hartzell GW 3rd, Stormo GD: Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci. CABIOS* 1990, 6:8192.
- [175] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993, 262:208214.

- [176] Gorodkin J, Heyer LJ, Brunak S, Stormo GD: Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci. CABIOS* 1997, 13:583586.
- [177] Ellrott K, Yang C, Sladek FM, Jiang T: Identifying transcription factor binding sites through Markov chain optimization. *Bioinforma. Oxf. Engl.* 2002, 18 Suppl 2:S100109.
- [178] Bulyk ML, Johnson PLF, Church GM: Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* 2002, 30:12551261.
- [179] Barash Y, Elidan G, Friedman N, Kaplan T?: <http://pluto.huji.ac.il/galelidan/papers/ElidanTFBS.pdf>.
- [180] Osada R, Zaslavsky E, Singh M: Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinforma. Oxf. Engl.* 2004, 20:35163525.
- [181] Tomovic A, Oakeley EJ: Position dependencies in transcription factor binding sites. *Bioinforma. Oxf. Engl.* 2007, 23:933941.
- [182] Owen GI, Zelent A: Origins and evolutionary diversification of the nuclear receptor superfamily. *Cell. Mol. Life Sci. CMLS* 2000, 57:809827.
- [183] Roulet E, Bucher P, Schneider R, Wingender E, Dusserre Y, Werner T, Mermod N: Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites. *J. Mol. Biol.* 2000, 297:833848.
- [184] Eddy Lab: Publications: Biological Sequence Analysis [<http://selab.janelia.org/cupbook.html>].
- [185] A tutorial on hidden Markov models and selected applications in speech recognition - Proceedings of the IEEE - [hmm\\_paper\\_rabiner.pdf](#).

- [186] Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouz Moreau Y: A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinforma. Oxf. Engl.* 2001, 17:11131122.
- [187] Jiang T, Wang L, Zhang K: Alignment of trees An alternative to tree edit. In *Comb. Pattern Matching.* edited by Crochemore M, Gusfield D Berlin, Heidelberg: Springer Berlin Heidelberg; 1994, 807:7586.
- [188] Kimura M, Ohta T: Probability of Gene Fixation in an Expanding Finite Population. *Proc. Natl. Acad. Sci.* 1974, 71:33773379.
- [189] Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol.* 1994, 2:2836.
- [190] The Analysis of Regulatory Sequences [<http://rsat.ulb.ac.be/course/index.html>].
- [191] Das MK, Dai H-K: A survey of DNA motif finding algorithms. *BMC Bioinformatics* 2007, 8:S21.
- [192] Rabiner LR: Rabiner, LR?: A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 1989, 77:257286.
- [193] Thakurta D.G, Stormo G.D : Finding regulatory elements in DNA sequence: <http://www.scionpublishing.com/shop/ProductImages/Bioinformatics%20Chapter%206.pdf>.
- [194] Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I: Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinforma. Oxf. Engl.* 2005, 21:26572666.
- [195] Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavese G, Pesole G, Rier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* 2005, 23:137144.

- [196] Tuerk C, Gold L: Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 1990, 249:505510.
- [197] Djordjevic M: SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. *Biomol. Eng.* 2007, 24:179189.
- [198] Liu X, Noll DM, Lieb JD, Clarke ND: DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res.* 2005, 15:421427.
- [199] Galas DJ, Schmitz A: DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* 1978, 5:31573170.
- [200] Fried M, Crothers DM: Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res.* 1981, 9:65056525.
- [201] Orlando V: Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem. Sci.* 2000, 25:99104.
- [202] Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA: Genome-wide location and function of DNA binding proteins. *Science* 2000, 290:23062309.
- [203] Johnson DS, Mortazavi A, Myers RM, Wold B: Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007, 316:14971502.
- [204] Kolchanov NA, Merkulova TI, Ignatieva EV, Ananko EA, Oshchepkov DY, Levitsky VG, Vasiliev GV, Klimova NV, Merkulov VM, Charles Hodgman T: Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes. *Brief. Bioinform.* 2007, 8:266274.
- [205] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov

- B, Saxel H, Kel AE, Wingender E: TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 2006, 34:D108110.
- [206] Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B: JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 2004, 32:D9194.
- [207] Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJM: ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinforma. Oxf. Engl.* 2006, 22:637640.
- [208] Portales-Casamar E, Arenillas D, Lim J, Swanson MI, Jiang S, McCallum A, Kirov S, Wasserman WW: The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res.* 2009, 37:D5460.
- [209] Jiang C, Xuan Z, Zhao F, Zhang MQ: TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res.* 2007, 35:D137140.
- [210] Kolchanov NA, Podkolodnaia OA, Ananko EA, Ignateva EV, Podkolodny? NL, Merkulov VM, Stepanenko IL, Pozdniakov MA, Belova OE, Grigorovich DA, Naumochkin AN: [Regulation of eukaryotic gene transcription: description in the TRRD database]. *Mol. Biol. (Mosk.)* 2001, 35:934942.
- [211] Ghosh D: Object-oriented transcription factors database (ooTFD). *Nucleic Acids Res.* 2000, 28:308310.
- [212] Sun H, Palaniswamy SK, Pohar TT, Jin VX, Huang TH-M, Davuluri RV: MPromDb: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-chip experimental data. *Nucleic Acids Res.* 2006, 34:D98103.

- [213] Ohler U, Liao G, Niemann H, Rubin GM: Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* 2002, 3:RESEARCH0087.
- [214] Gershenzon NI, Trifonov EN, Ioshikhes IP: The features of *Drosophila* core promoters revealed by statistical analysis. *BMC Genomics* 2006, 7:161.
- [215] modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezhikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, Booth BW, Brooks AN, Dai Q, Davis CA, Duff MO, Feng X, Gorchakov AA, Gu T, Henikoff JG, Kapranov P, Li R, MacAlpine HK, Malone J, Minoda A, Nordman J, Okamura K, Perry M, Powell SK, Riddle NC, Sakai A, Samsonova A, Sandler JE, Schwartz YB, Sher N, Spokony R, Sturgill D, van Baren M, Wan KH, Yang L, Yu C, Feingold E, Good P, Guyer M, Lowdon R, Ahmad K, Andrews J, Berger B, Brenner SE, Brent MR, Cherbas L, Elgin SCR, Gingeras TR, Grossman R, Hoskins RA, Kaufman TC, Kent W, Kuroda MI, Orr-Weaver T, Perrimon N, Pirrotta V, Posakony JW, Ren B, Russell S, Cherbas P, Graveley BR, Lewis S, Micklem G, Oliver B, Park PJ, Celniker SE, Henikoff S, Karpen GH, Lai EC, MacAlpine DM, Stein LD, White KP, Kellis M: Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 2010, 330:17871797.
- [216] Guertin MJ, Lis JT: Chromatin landscape dictates HSF binding to target DNA elements. *PLoS Genet.* 2010, 6.
- [217] Eaton ML, Prinz JA, MacAlpine HK, Tretyakov G, Kharchenko PV, MacAlpine DM: Chromatin signatures of the *Drosophila* replication program. *Genome Res.* 2011, 21:164174.
- [218] Berezhikov E, Robine N, Samsonova A, Westholm JO, Naqvi A, Hung J-H, Okamura K, Dai Q, Bortolamiol-Becet D, Martin R, Zhao Y, Zamore PD, Hannon

- GJ, Marra MA, Weng Z, Perrimon N, Lai EC: Deep annotation of *Drosophila melanogaster* microRNAs yields insights into their processing, modification, and emergence. *Genome Res.* 2011, 21:203215.
- [219] Paro S, Li X, OConnell MA, Keegan LP: Regulation and functions of ADAR in *Drosophila*. *Curr. Top. Microbiol. Immunol.* 2012, 353:221236.
- [220] Ne N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony RF, Chen J, Hwang L, Cheng C, Auburn RP, Davis MB, Domanus M, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Victorsen A, Bild NA, Grundstad AJ, Hanley D, MacAlpine DM, Mannervik M, Venken K, Bellen H, White R, Gerstein M, Russell S, Grossman RL, Ren B, Posakony JW, Kellis M, White KP: A cis-regulatory map of the *Drosophila* genome. *Nature* 2011, 471:527531.
- [221] modMine: Home [<http://intermine.modencode.org/>].
- [222] Contrino S, Smith RN, Butano D, Carr A, Hu F, Lyne R, Rutherford K, Kalderimis A, Sullivan J, Carbon S, Kephart ET, Lloyd P, Stinson EO, Washington NL, Perry MD, Ruzanov P, Zha Z, Lewis SE, Stein LD, Micklem G: modMine: flexible access to modENCODE data. *Nucleic Acids Res.* 2012, 40:D10821088.
- [223] DBTSS HOME [[http://dbtss\\_old.hgc.jp/hg18\\_mm9\\_3/](http://dbtss_old.hgc.jp/hg18_mm9_3/)].
- [224] Bentley DR: Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* 2006, 16:545552.
- [225] Yamashita R, Sathira NP, Kanai A, Tanimoto K, Arauchi T, Tanaka Y, Hashimoto S-I, Sugano S, Nakai K, Suzuki Y: Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res.* 2011, 21:775789.
- [226] Ohler U: Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction. *Nucleic Acids Res.* 2006, 34:59435950.

- [227] Zenzie-Gregory B, Khachi A, Garraway IP, Smale ST: Mechanism of initiator-mediated transcription: evidence for a functional interaction between the TATA-binding protein and DNA in the absence of a specific recognition sequence. *Mol. Cell. Biol.* 1993, 13:38413849.
- [228] Martinez E, Zhou Q, L'Etoile ND, Oelgeschlör T, Berk AJ, Roeder RG: Core promoter-specific function of a mutant transcription factor TFIID defective in TATA-box binding. *Proc. Natl. Acad. Sci. U. S. A.* 1995, 92:1186411868.
- [229] Tsai FT, Sigler PB: Structural basis of preinitiation complex assembly on human pol II promoters. *Embo J.* 2000, 19:2536.
- [230] O'Shea-Greenfield A, Smale ST: Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. *J. Biol. Chem.* 1992, 267:13911402.
- [231] Emami KH, Jain A, Smale ST: Mechanism of synergy between TATA and initiator: synergistic binding of TFIID following a putative TFIIA-induced isomerization. *Genes Dev.* 1997, 11:30073019.
- [232] Zhou T, Chiang CM: The intronless and TATA-less human TAF(II)55 gene contains a functional initiator and a downstream promoter element. *J. Biol. Chem.* 2001, 276:2550325511.
- [233] Lee D-H, Gershenzon N, Gupta M, Ioshikhes IP, Reinberg D, Lewis BA: Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Mol. Cell. Biol.* 2005, 25:96749686.
- [234] Gershenzon NI, Ioshikhes IP: Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinforma. Oxf. Engl.* 2005, 21:12951300.
- [235] SRA002054:[<http://trace.ddbj.nig.ac.jp/DRAsearch/submission?acc=SRA002054>].
- [236] FASTX-Toolkit [[http://hannonlab.cshl.edu/fastx\\_toolkit/commandline.html](http://hannonlab.cshl.edu/fastx_toolkit/commandline.html)].



- [237] Novocraft.com: Novocraft [<http://www.novocraft.com/main/index.php>].
- [238] Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* 2010, 26:841842.
- [239] Stein LD: Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief. Bioinform.* 2013, 14:162171.
- [240] Suzuki Y, Taira H, Tsunoda T, Mizushima-Sugano J, Sese J, Hata H, Ota T, Isogai T, Tanaka T, Morishita S, Okubo K, Sakaki Y, Nakamura Y, Suyama A, Sugano S: Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.* 2001, 2:388393.
- [241] Carninci P: Tagging mammalian transcription complexity. *Trends Genet. Tig* 2006, 22:501510.
- [242] Zhao X, Valen E, Parker BJ, Sandelin A: Systematic clustering of transcription start site landscapes. *Plos One* 2011, 6:e23409.
- [243] Dreos R, Ambrosini G, Cavin Perier R, Bucher P: EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.* 2012, 41:D157D164.
- [244] Bryne JC, Valen E, Tang M-HE, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A: JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* 2008, 36:D102106.
- [245] Turatsinze J-V, Thomas-Chollier M, Defrance M, van Helden J: Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc.* 2008, 3:15781588.
- [246] Van Helden J: Regulatory sequence analysis tools. *Nucleic Acids Res.* 2003, 31:35933596.

- [247] Keich U, Pevzner PA: Subtle motifs: defining the limits of motif finding algorithms. *Bioinforma. Oxf. Engl.* 2002, 18:13821390.
- [248] Frith MC, Li MC, Weng Z: Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* 2003, 31:36663668.
- [249] Bailey TL, Elkan C: The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol. Ismb Int. Conf. Intell. Syst. Mol. Biol.* 1995, 3:2129.
- [250] Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: Quantifying similarity between motifs. *Genome Biol.* 2007, 8:R24.
- [251] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 2000, 25:2529.
- [252] Corden J, Wasylyk B, Buchwalder A, Sassone-Corsi P, Kedinger C, Chambon P: Promoter sequences of eukaryotic protein-coding genes. *Science* 1980, 209:14061414.
- [253] Merzendorfer H, Zimoch L: Chitin metabolism in insects: structure, function and regulation of chitin synthases and chitinases. *J. Exp. Biol.* 2003, 206:43934412.
- [254] Mito Y, Henikoff JG, Henikoff S: Genome-scale profiling of histone H3.3 replacement patterns. *Nat. Genet.* 2005, 37:10901097.
- [255] Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 2005, 309:626630.

- [256] Illingworth RS, Bird AP: CpG islands—a rough guide. *Febs Lett.* 2009, 583:17131720.
- [257] Mohn F, Schbeler D: Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends Genet.* *Tig* 2009, 25:129136.
- [258] Valen E, Sandelin A: Genomic and chromatin signals underlying transcription start-site selection. *Trends Genet.* *Tig* 2011, 27:475485.
- [259] Kutach AK, Kadonaga JT: The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol. Cell. Biol.* 2000, 20:47544764.
- [260] Bajic VB, Tan SL, Christoffels A, Schnbach C, Lipovich L, Yang L, Hofmann O, Kruger A, Hide W, Kai C, Kawai J, Hume DA, Carninci P, Hayashizaki Y: Mice and men: their promoter properties. *Plos Genet.* 2006, 2:e54.
- [261] Theisen JWM, Lim CY, Kadonaga JT: Three key subregions contribute to the function of the downstream RNA polymerase II core promoter. *Mol. Cell. Biol.* 2010, 30:34713479.
- [262] Anish R, Hossain MB, Jacobson RH, Takada S: Characterization of Transcription from TATA-Less Promoters: Identification of a New Core Promoter Element XCPE2 and Analysis of Factor Requirements. *Plos One* 2009, 4:e5103.
- [263] Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E: Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* 2007, 389:5265.
- [264] Guillon N, Tirode F, Boeva V, Zynovyev A, Barillot E, Delattre O: The oncogenic EWS-FLI1 protein binds in vivo GGAA microsatellite sequences with potential transcriptional activation function. *Plos One* 2009, 4:e4932.
- [265] Cohen CJ, Lock WM, Mager DL: Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* 2009, 448:105114.

- [266] Lee T-H, Maheshri N: A regulatory role for repeated decoy transcription factor binding sites in target gene expression. *Mol. Syst. Biol.* 2012, 8:576.
- [267] Glazov EA, Pheasant M, Nahkuri S, Mattick JS: Evidence for control of splicing by alternative RNA secondary structures in Dipteran homothorax pre-mRNA. *Rna Biol.* 2006, 3:3639.
- [268] Li M, Wen S, Guo X, Bai B, Gong Z, Liu X, Wang Y, Zhou Y, Chen X, Liu L, Chen R: The novel long non-coding RNA CRG regulates *Drosophila* locomotor behavior. *Nucleic Acids Res.* 2012, 40:1171411727.
- [269] Gokhale K, Patil DP, Dhotre DP, Dixit R, Mendki MJ, Patole MS, Shouche YS: Transcriptome analysis of *Anopheles stephensi* embryo using expressed sequence tags. *J. Biosci.* 2013, 38:301309.
- [270] Lim L, Manser E, Leung T, Hall C: Regulation of phosphorylation pathways by p21 GTPases. The p21 Ras-related Rho subfamily and its role in phosphorylation signalling pathways. *Eur. J. Biochem. Febs* 1996, 242:171185.
- [271] Van Aelst L, DSouza-Schorey C: Rho GTPases and signaling networks. *Genes Dev.* 1997, 11:22952322.
- [272] Harden N, Ricos M, Ong YM, Chia W, Lim L: Participation of small GTPases in dorsal closure of the *Drosophila* embryo: distinct roles for Rho subfamily proteins in epithelial morphogenesis. *J. Cell Sci.* 1999, 112 ( Pt 3):273284.
- [273] Guo X, Macleod GT, Wellington A, Hu F, Panchumarthi S, Schoenfield M, Marin L, Charlton MP, Atwood HL, Zinsmaier KE: The GTPase dMiro is required for axonal transport of mitochondria to *Drosophila* synapses. *Neuron* 2005, 47:379393.
- [274] Kamiyama D, Chiba A: Endogenous activation patterns of Cdc42 GTPase within *Drosophila* embryos. *Science* 2009, 324:13381340.

- [275] Zhao L, Becnel JJ, Clark GG, Linthicum KJ: Expression of AeaHsp26 and AeaHsp83 in *Aedes aegypti* (Diptera: Culicidae) larvae and pupae in response to heat shock stress. *J. Med. Entomol.* 2010, 47:367375.
- [276] Baena-L LA, Alonso J, Rodriguez J, SantarF: The expression of heat shock protein HSP60A reveals a dynamic mitochondrial pattern in *Drosophila melanogaster* embryos. *J. Proteome Res.* 2008, 7:27802788.
- [277] Hoffmann JA: The immune response of *Drosophila*. *Nature* 2003, 426:3338.
- [278] Wilson R, Chen C, Ratcliffe NA: Innate immunity in insects: the role of multiple, endogenous serum lectins in the recognition of foreign invaders in the cockroach, *Blaberus discoidalis*. *J. Immunol. Baltim. Md 1950* 1999, 162:15901596.
- [279] Yu X-Q, Zhu Y-F, Ma C, Fabrick JA, Kanost MR: Pattern recognition proteins in *Manduca sexta* plasma. *Insect Biochem. Mol. Biol.* 2002, 32:12871293.
- [280] Holmskov U, Thiel S, Jensenius JC: Collections and ficolins: humoral lectins of the innate immune defense. *Annu. Rev. Immunol.* 2003, 21:547578.
- [281] Takeda K, Kaisho T, Akira S: Toll-like receptors. *Annu. Rev. Immunol.* 2003, 21:335376.
- [282] Gobert V, Gottar M, Matskevich AA, Rutschmann S, Royet J, Belvin M, Hoffmann JA, Ferrandon D: Dual activation of the *Drosophila* toll pathway by two pattern recognition receptors. *Science* 2003, 302:21262130.
- [283] Michel T, Reichhart JM, Hoffmann JA, Royet J: *Drosophila* Toll is activated by Gram-positive bacteria through a circulating peptidoglycan recognition protein. *Nature* 2001, 414:756759.
- [284] Choe K-M, Werner T, Stven S, Hultmark D, Anderson KV: Requirement for a peptidoglycan recognition protein (PGRP) in Relish activation and antibacterial immune responses in *Drosophila*. *Science* 2002, 296:359362.

- [285] Gottar M, Gobert V, Michel T, Belvin M, Duyk G, Hoffmann JA, Ferrandon D, Royet J: The *Drosophila* immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein. *Nature* 2002, 416:640644.
- [286] Rt M, Manfrulli P, Pearson A, Mathey-Prevot B, Ezekowitz RAB: Functional genomic analysis of phagocytosis and identification of a *Drosophila* receptor for *E. coli*. *Nature* 2002, 416:644648.
- [287] Takehana A, Katsuyama T, Yano T, Oshima Y, Takada H, Aigaki T, Kurata S: Overexpression of a pattern-recognition receptor, peptidoglycan-recognition protein-LE, activates imd/relish-mediated antibacterial defense and the phenoloxidase cascade in *Drosophila* larvae. *Proc. Natl. Acad. Sci. U. S. A.* 2002, 99:1370513710.
- [288] Weiss BL, Wu Y, Schwank JJ, Tolwinski NS, Aksoy S: An insect symbiosis is influenced by bacterium-specific polymorphisms in outer-membrane protein A. *Proc. Natl. Acad. Sci. U. S. A.* 2008, 105:1508815093.
- [289] Wang J, Wu Y, Yang G, Aksoy S: Interactions between mutualist *Wigglesworthia* and tsetse peptidoglycan recognition protein (PGRP-LB) influence trypanosome transmission. *Proc. Natl. Acad. Sci. U. S. A.* 2009, 106:1213312138.
- [290] Lemaitre B, Hoffmann J: The host defense of *Drosophila melanogaster*. *Annu. Rev. Immunol.* 2007, 25:697743.
- [291] Buchon N, Broderick NA, Chakrabarti S, Lemaitre B: Invasive and indigenous microbiota impact intestinal stem cell activity through multiple pathways in *Drosophila*. *Genes Dev.* 2009, 23:23332344.
- [292] Dostert C, Jouanguy E, Irving P, Troxler L, Galiana-Arnoux D, Hetru C, Hoffmann JA, Imler J-L: The Jak-STAT signaling pathway is required but not sufficient for the antiviral response of *drosophila*. *Nat. Immunol.* 2005, 6:946953.
- [293] IID [[http://bordensteinlab.vanderbilt.edu/IIDtest\\_immunity.php](http://bordensteinlab.vanderbilt.edu/IIDtest_immunity.php)].

- [294] Hao Z, Kasumba I, Aksoy S: Proventriculus (cardia) plays a crucial role in immunity in tsetse fly (Diptera: Glossinidae). *Insect Biochem. Mol. Biol.* 2003, 33:11551164.
- [295] Nayduch D, Aksoy S: Refractoriness in tsetse flies (Diptera: Glossinidae) may be a matter of timing. *J. Med. Entomol.* 2007, 44:660665.
- [296] Levashina EA, Moita LF, Blandin S, Vriend G, Lagueux M, Kafatos FC: Conserved role of a complement-like protein in phagocytosis revealed by dsRNA knockout in cultured cells of the mosquito, *Anopheles gambiae*. *Cell* 2001, 104:709718.
- [297] Gtz P: Encapsulation in Arthropods. In *Immun. Invertebr.* edited by Brehn M Berlin, Heidelberg: Springer Berlin Heidelberg; 1986:153170.
- [298] Lanot R, Zachary D, Holder F, Meister M: Postembryonic hematopoiesis in *Drosophila*. *Dev. Biol.* 2001, 230:243257.
- [299] Kumar S, Christophides GK, Cantera R, Charles B, Han YS, Meister S, Dimopoulos G, Kafatos FC, Barillas-Mury C: The role of reactive oxygen species on *Plasmodium melanotic* encapsulation in *Anopheles gambiae*. *Proc. Natl. Acad. Sci. U. S. A.* 2003, 100:1413914144.
- [300] Vallet-Gely I, Lemaitre B, Boccard F: Bacterial strategies to overcome insect defences. *Nat. Rev. Microbiol.* 2008, 6:302313.
- [301] Munks RJL, SantAnna MRV, Grail W, Gibson W, Igglesden T, Yoshiyama M, Lehane SM, Lehane MJ: Antioxidant gene expression in the blood-feeding fly *Glossina morsitans morsitans*. *Insect Mol. Biol.* 2005, 14:483491.
- [302] Rhamhul N.U, PhD thesis: [http://research-archive.liv.ac.uk/6053/4/RamphulUrv\\_Feb2012\\_6053.pdf](http://research-archive.liv.ac.uk/6053/4/RamphulUrv_Feb2012_6053.pdf).
- [303] Meredith JM, Munks RJL, Grail W, Hurd H, Eggleston P, Lehane MJ: A novel association between clustered NF-kappaB and C/EBP binding sites is

required for immune regulation of mosquito Defensin genes. *Insect Mol. Biol.* 2006, 15:393401.

- [304] Hernandez-Romano J, Carlos-Rivera FJ, Salgado H, Lamadrid-Figueroa H, Valverde-Gardu Rodriguez MH, Martinez-Barnette J: Immunity related genes in dipterans share common enrichment of AT-rich motifs in their 5 regulatory regions that are potentially involved in nucleosome formation. *BMC Genomics* 2008, 9:326.
- [305] Sieglaff DH, Dunn WA, Xie XS, Megy K, Marinotti O, James AA: Comparative genomics allows the discovery of cis-regulatory elements in mosquitoes. *Proc. Natl. Acad. Sci. U. S. A.* 2009, 106:30533058.
- [306] Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, Barillas-Mury C, Bian G, Blandin S, Christensen BM, Dong Y, Jiang H, Kanost MR, Koutsos AC, Levashina EA, Li J, Ligoxygakis P, Maccallum RM, Mayhew GF, Mendes A, Michel K, Osta MA, Paskewitz S, Shin SW, Vlachou D, Wang L, Wei W, Zheng L, Zou Z, Severson DW, Raikhel AS, Kafatos FC, Dimopoulos G, Zdobnov EM, Christophides GK: Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* 2007, 316:17381743.
- [307] Kasprzyk A: BioMart: driving a paradigm change in biological data management. *Database J. Biol. Databases Curation* 2011, 2011:bar049.
- [308] Kersey PJ, Staines DM, Lawson D, Kulesha E, Derwent P, Humphrey JC, Hughes DST, Keenan S, Kerhornou A, Koscielny G, Langridge N, McDowall MD, Megy K, Maheswari U, Nuhn M, Paulini M, Pedro H, Toneva I, Wilson D, Yates A, Birney E: Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.* 2012, 40:D9197.
- [309] Li L, Stoeckert CJ Jr, Roos DS: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003, 13:21782189.



- [310] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J. Mol. Biol.* 1990, 215:403410.
- [311] Elkon R: Genome-Wide In Silico Identification of Transcriptional Regulators Controlling the Cell Cycle in Human Cells. *Genome Res.* 2003, 13:773780.
- [312] Cooper SJ: Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* 2005, 16:110.
- [313] Tabach Y, Brosh R, Buggan Y, Reiner A, Zuk O, Yitzhaky A, Koudritsky M, Rotter V, Domany E: Wide-Scale Analysis of Human Functional Transcription Factor Binding Reveals a Strong Bias towards the Transcription Start Site. *PLoS ONE* 2007, 2:e807.
- [314] Koudritsky M, Domany E: Positional distribution of human transcription factor binding sites. *Nucleic Acids Res.* 2008, 36:67956805.
- [315] Ao W, Gaudet J, Kent WJ, Muttumu S, Mango SE: Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* 2004, 305:17431746.
- [316] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 2009, 37:W202208.
- [317] Defrance M, van Helden J: info-gibbs: a motif discovery algorithm that directly optimizes information content during sampling. *Bioinforma. Oxf. Engl.* 2009, 25:27152722.
- [318] Mahony S, Benos PV: STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* 2007, 35:W253258.
- [319] Biggin MD, McGinnis W: Regulation of segmentation and segmental identity by *Drosophila* homeoproteins: the role of DNA binding in functional activity and specificity. *Dev. Camb. Engl.* 1997, 124:44254433.

- [320] Senger K, Armstrong GW, Rowell WJ, Kwan JM, Markstein M, Levine M: Immunity regulatory DNAs share common organizational features in *Drosophila*. *Mol. Cell* 2004, 13:1932.
- [321] Senger K, Harris K, Levine M: GATA factors participate in tissue-specific immune responses in *Drosophila* larvae. *Proc. Natl. Acad. Sci. U. S. A.* 2006, 103:1595715962.
- [322] Uvell H, Engström Y: A multilayered defense against infection: combinatorial control of insect immune genes. *Trends Genet. TIG* 2007, 23:342349.
- [323] Tanji T, Hu X, Weber ANR, Ip YT: Toll and IMD pathways synergistically activate an innate immune response in *Drosophila melanogaster*. *Mol. Cell. Biol.* 2007, 27:45784588.
- [324] Han S-H, Ryu J-H, Oh C-T, Nam K-B, Nam H-J, Jang I-H, Brey PT, Lee W-J: The moleskin gene product is essential for Caudal-mediated constitutive antifungal Drosomycin gene expression in *Drosophila* epithelia. *Insect Mol. Biol.* 2004, 13:323327.
- [325] Mehrpour M, Botti J, Codogno P: Mechanisms and regulation of autophagy in mammalian cells. *Atlas Genet. Cytogenet. Oncol. Haematol.* 2012.
- [326] Ma D, Lin JD: Circadian regulation of autophagy rhythm through transcription factor C/EBP?. *Autophagy* 2012, 8:124125.
- [327] Li T, Lu L: Functional role of CCCTC binding factor (CTCF) in stress-induced apoptosis. *Exp. Cell Res.* 2007, 313:30573065.
- [328] Kim Y-I, Ryu T, Lee J, Heo Y-S, Ahnn J, Lee S-J, Yoo O: A genetic screen for modifiers of *Drosophila* caspase Dcp-1 reveals caspase involvement in autophagy and novel caspase-related genes. *BMC Cell Biol.* 2010, 11:9.
- [329] Hew HC, Liu H, Miki Y, Yoshida K: PKC $\beta$  regulates Mdm2 independently of p53 in the apoptotic response to DNA damage. *Mol. Carcinog.* 2011, 50:719731.

- [330] Pang S, Chen D, Zhang A, Qin X, Yan B: Genetic analysis of the LAMP-2 gene promoter in patients with sporadic Parkinsons disease. *Neurosci. Lett.* 2012, 526:6367.
- [331] Inami Y, Waguri S, Sakamoto A, Kouno T, Nakada K, Hino O, Watanabe S, Ando J, Iwadate M, Yamamoto M, Lee M-S, Tanaka K, Komatsu M: Persistent activation of Nrf2 through p62 in hepatocellular carcinoma cells. *J. Cell Biol.* 2011, 193:275284.
- [332] Guan J-L, Simon AK, Prescott M, Menendez JA, Liu F, Wang F, Wang C, Wolvetang E, Vazquez-Martin A, Zhang J: Autophagy in stem cells. *Autophagy* 2013, 9:830849.
- [333] Pietrocola F, Izzo V, Niso-Santano M, Vacchelli E, Galluzzi L, Maiuri MC, Kroemer G: Regulation of autophagy by stress-responsive transcription factors. *Semin. Cancer Biol.* 2013.
- [334] Chen Z-H, Kim HP, Scieurba FC, Lee S-J, Feghali-Bostwick C, Stolz DB, Dhir R, Landreneau RJ, Schuchert MJ, Yousem SA, Nakahira K, Pilewski JM, Lee JS, Zhang Y, Ryter SW, Choi AMK: Egr-1 regulates autophagy in cigarette smoke-induced chronic obstructive pulmonary disease. *PloS One* 2008, 3:e3316.
- [335] Van der Stap L, MSc Thesis: [http://sdsu-dspace.calstate.edu/bitstream/handle/10211.10/1954/VanDerStap\\_Lena.pdf?sequence=1](http://sdsu-dspace.calstate.edu/bitstream/handle/10211.10/1954/VanDerStap_Lena.pdf?sequence=1).
- [336] Hornung V, Ablasser A, Charrel-Dennis M, Bauernfeind F, Horvath G, Caffrey DR, Latz E, Fitzgerald KA: AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. *Nature* 2009, 458:514518.
- [337] Sabbagh L, Bourbonni M, Denis F, Sly R-P: Cloning and functional characterization of the murine caspase-3 gene promoter. *DNA Cell Biol.* 2006, 25:104115.
- [338] Liu Y, Ringn: Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis. *Genome Biol.* 2007, 8:R77.

- [339] Liu W, Wang G, Yakovlev AG: Identification and functional analysis of the rat caspase-3 gene promoter. *J. Biol. Chem.* 2002, 277:82738278.
- [340] Nenoï M, Ichimura S, Mita K, Yukawa O, Cartwright IL: Regulation of the catalase gene promoter by Sp1, CCAAT-recognizing factors, and a WT1/Egr-related factor in hydrogen peroxide-resistant HP100 cells. *Cancer Res.* 2001, 61:58855894.
- [341] Rodrigues J, Agrawal N, Sharma A, Malhotra P, Adak T, Chauhan VS, Bhatnagar RK: Transcriptional analysis of an immune-responsive serine protease from Indian malarial vector, *Anopheles culicifacies*. *BMC Mol. Biol.* 2007, 8:33.
- [342] Zou Z, Shin SW, Alvarez KS, Bian G, Kokoza V, Raikhel AS: Mosquito RUNX4 in the immune regulation of PPO gene expression and its effect on avian malaria parasite infection. *Proc. Natl. Acad. Sci. U. S. A.* 2008, 105:1845418459.
- [343] Raimond J, Rouleux F, Monsigny M, Legrand A: The second intron of the human galectin-3 gene has a strong promoter activity down-regulated by p53. *FEBS Lett.* 1995, 363:165169.
- [344] Erl W, Hansson GK, de Martin R, Draude G, Weber KS, Weber C: Nuclear factor-kappa B regulates induction of apoptosis and inhibitor of apoptosis protein-1 expression in vascular smooth muscle cells. *Circ. Res.* 1999, 84:668677.
- [345] Betz A, Ryoo HD, Steller H, Darnell JE Jr: STAT92E is a positive regulator of *Drosophila* inhibitor of apoptosis 1 (DIAP/1) and protects against radiation-induced apoptosis. *Proc. Natl. Acad. Sci. U. S. A.* 2008, 105:1380513810.
- [346] Dong Z, Nishiyama J, Yi X, Venkatachalam MA, Denton M, Gu S, Li S, Qiang M: Gene promoter of apoptosis inhibitory protein IAP2: identification of enhancer elements and activation by severe hypoxia. *Biochem. J.* 2002, 364:413421.
- [347] Takeda K, Akira S: Toll receptors and pathogen resistance. *Cell. Microbiol.* 2003, 5:143153.

- [348] Meier D, Schindler D: Fanconi anemia core complex gene promoters harbor conserved transcription regulatory elements. *PloS One* 2011, 6:e22911.
- [349] Horvai A, Palinski W, Wu H, Moulton KS, Kalla K, Glass CK: Scavenger receptor A gene regulatory elements target gene expression to macrophages and to foam cells of atherosclerotic lesions. *Proc. Natl. Acad. Sci. U. S. A.* 1995, 92:53915395.
- [350] Danielli A, Kafatos FC, Loukeris TG: Cloning and characterization of four *Anopheles gambiae* serpin isoforms, differentially induced in the midgut by *Plasmodium berghei* invasion. *J. Biol. Chem.* 2003, 278:41844193.
- [351] Zhang M, Magit D, Sager R: Expression of maspin in prostate cells is regulated by a positive ets element and a negative hormonal responsive element site recognized by androgen receptor. *Proc. Natl. Acad. Sci. U. S. A.* 1997, 94:56735678.
- [352] Park DS, Shin SW, Hong SD, Park HY: Immunological detection of serpin in the fall webworm, *Hyphantria cunea* and its inhibitory activity on the prophe-noloxidase system. *Mol. Cells* 2000, 10:186192.
- [353] Kalsheker N, Morley S, Morgan K: Gene regulation of the serine proteinase inhibitors alpha1-antitrypsin and alpha1-antichymotrypsin. *Biochem. Soc. Trans.* 2002, 30:9398.
- [354] Koziczak M, Krek W, Nagamine Y: Pocket protein-independent repression of urokinase-type plasminogen activator and plasminogen activator inhibitor 1 gene expression by E2F1. *Mol. Cell. Biol.* 2000, 20:20142022.
- [355] Bailey CM, Khalkhali-Ellis Z, Kondo S, Margaryan NV, Seftor REB, Wheaton WW, Amir S, Pins MR, Schutte BC, Hendrix MJC: Mammary serine protease inhibitor (Maspin) binds directly to interferon regulatory factor 6: identification of a novel serpin partnership. *J. Biol. Chem.* 2005, 280:3421034217.
- [356] Boyapati A, Ren B, Zhang D-E: SERPINB13 is a novel RUNX1 target gene. *Biochem. Biophys. Res. Commun.* 2011, 411:115120.

- [357] Woenckhaus M, Bubendorf L, Dalquen P, Foerster J, Blaszyk H, Mirlacher M, Soler M, Dietmaier W, Sauter G, Hartmann A, Wild PJ: Nuclear and cytoplasmic Maspin expression in primary non-small cell lung cancer. *J. Clin. Pathol.* 2006, 60:483486.
- [358] Liao H, Hyman MC, Lawrence DA, Pinsky DJ: Molecular regulation of the PAI-1 gene by hypoxia: contributions of Egr-1, HIF-1alpha, and C/EBPalpha. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* 2007, 21:935949.
- [359] Campbell CL, Black WC, Hess AM, Foy BD: Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. *BMC Genomics* 2008, 9:425.
- [360] Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S, Blass C, Brey PT, Collins FH, Danielli A, Dimopoulos G, Hetru C, Hoa NT, Hoffmann JA, Kanzok SM, Letunic I, Levashina EA, Loukeris TG, Lycett G, Meister S, Michel K, Moita LF, Mller H-M, Osta MA, Paskewitz SM, Reichhart J-M, Rzhetsky A, Troxler L, Vernick KD, Vlachou D, Volz J, von Mering C, Xu J, Zheng L, Bork P, Kafatos FC: Immunity-related genes and gene families in *Anopheles gambiae*. *Science* 2002, 298:159165.
- [361] Zou Z, Evans JD, Lu Z, Zhao P, Williams M, Sumathipala N, Hetru C, Hultmark D, Jiang H: Comparative genomic analysis of the *Tribolium* immune system. *Genome Biol.* 2007, 8:R177.
- [362] Winter F, Edaye S, Huttenhofer A, Brunel C: *Anopheles gambiae* miRNAs as actors of defence reaction against *Plasmodium* invasion. *Nucleic Acids Res.* 2007, 35:69536962.
- [363] Bonizzoni M, Dunn W, Campbell CL, Olson KE, Dimon MT, Marinotti O, James AA: RNA-seq analyses of blood-induced changes in gene expression in the mosquito vector species, *Aedes aegypti*. *BMC Genomics* 2011, 12:82.

- [364] Skalsky RL, Vanlandingham DL, Scholle F, Higgs S, Cullen BR: Identification of microRNAs expressed in two mosquito vectors, *Aedes albopictus* and *Culex quinquefasciatus*. *BMC Genomics* 2010, 11:119.
- [365] Maiuri MC, Zalckvar E, Kimchi A, Kroemer G: Self-eating and self-killing: crosstalk between autophagy and apoptosis. *Nat. Rev. Mol. Cell Biol.* 2007, 8:741752.
- [366] Berry DL, Baehrecke EH: Growth arrest and autophagy are required for salivary gland cell degradation in *Drosophila*. *Cell* 2007, 131:11371148.
- [367] Denton D, Shrivage B, Simin R, Mills K, Berry DL, Baehrecke EH, Kumar S: Autophagy, not apoptosis, is essential for midgut cell death in *Drosophila*. *Curr. Biol. CB* 2009, 19:17411746.
- [368] Klichko VI, Radyuk SN, Orr WC: Profiling catalase gene expression in *Drosophila melanogaster* during development and aging. *Arch. Insect Biochem. Physiol.* 2004, 56:3450.
- [369] Binari R, Perrimon N: Stripe-specific regulation of pair-rule genes by hopscotch, a putative Jak family tyrosine kinase in *Drosophila*. *Genes Dev.* 1994, 8:300312.
- [370] Halfon MS, Hashimoto C, Keshishian H: The *Drosophila* toll gene functions zygotically and is necessary for proper motoneuron and muscle development. *Dev. Biol.* 1995, 169:151167.
- [371] Belvin MP, Anderson KV: A conserved signaling pathway: the *Drosophila* toll-dorsal pathway. *Annu. Rev. Cell Dev. Biol.* 1996, 12:393416.
- [372] Qiu P, Pan PC, Govind S: A role for the *Drosophila* Toll/Cactus pathway in larval hematopoiesis. *Dev. Camb. Engl.* 1998, 125:19091920.
- [373] Agaisse H, Perrimon N: The roles of JAK/STAT signaling in *Drosophila* immune responses. *Immunol. Rev.* 2004, 198:7282.

- [374] Bina S, PhD Thesis: <http://ediss.uni-goettingen.de/bitstream/handle/11858/00-1735-0000-0006-B4FF-E/bina.pdf?sequence=1>.
- [375] Nsslein-Volhard C, Wieschaus E: Mutations affecting segment number and polarity in *Drosophila*. *Nature* 1980, 287:795801.
- [376] Wasserman SA: A conserved signal transduction pathway regulating the activity of the rel-like proteins dorsal and NF-kappa B. *Mol. Biol. Cell* 1993, 4:767771.
- [377] Lemaitre B, Nicolas E, Michaut L, Reichhart JM, Hoffmann JA: The dorsoventral regulatory gene cassette *sple/Toll/cactus* controls the potent antifungal response in *Drosophila* adults. *Cell* 1996, 86:973983.
- [378] Hoffmann JA, Reichhart J-M: *Drosophila* innate immunity: an evolutionary perspective. *Nat. Immunol.* 2002, 3:121126.
- [379] Rushlow C: Dorsoventral patterning: a serpin pinned down at last. *Curr. Biol. CB* 2004, 14:R1618.
- [380] Hashimoto C, Kim DR, Weiss LA, Miller JW, Morisato D: Spatial regulation of developmental signaling by a serpin. *Dev. Cell* 2003, 5:945950.
- [381] Flynt AS, Lai EC: Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nat. Rev. Genet.* 2008, 9:831842.
- [382] Junell A, Uvell H, Davis MM, Edlundh-Rose E, Antonsson A, Pick L, Engström Y: The POU transcription factor *Drifter/Ventral veinless* regulates expression of *Drosophila* immune defense genes. *Mol. Cell. Biol.* 2010, 30:36723684.
- [383] Ryu J-H, Kim S-H, Lee H-Y, Bai JY, Nam Y-D, Bae J-W, Lee DG, Shin SC, Ha E-M, Lee W-J: Innate immune homeostasis by the homeobox gene *caudal* and commensal-gut mutualism in *Drosophila*. *Science* 2008, 319:777782.
- [384] Clayton AM, Cirimotich CM, Dong Y, Dimopoulos G: *Caudal* is a negative regulator of the *Anopheles* IMD pathway that controls resistance to *Plasmodium falciparum* infection. *Dev. Comp. Immunol.* 2013, 39:323332.



- [385] Gao L, Sun C, Qiu H-L, Liu H, Shao H-J, Wang J, Li W-X: Cloning and characterization of a novel human zinc finger gene, hKid3, from a C2H2-ZNF enriched human embryonic cDNA library. *Biochem. Biophys. Res. Commun.* 2004, 325:11451152.
- [386] Fujita T, Miyamoto M, Kimura Y, Hammer J, Taniguchi T: Involvement of a cis-element that binds an H2TF-1/NF kappa B like factor(s) in the virus-induced interferon-beta gene expression. *Nucleic Acids Res.* 1989, 17:33353346.
- [387] Tanaka N, Kawakami T, Taniguchi T: Recognition DNA sequences of interferon regulatory factor 1 (IRF-1) and IRF-2, regulators of cell growth and the interferon system. *Mol. Cell. Biol.* 1993, 13:45314538.
- [388] Kumar A, Kumar A, Michael P, Brabant D, Parissenti AM, Ramana CV, Xu X, Parrillo JE: Human serum from patients with septic shock activates transcription factors STAT1, IRF1, and NF-kappaB and induces apoptosis in human cardiac myocytes. *J. Biol. Chem.* 2005, 280:4261942626.
- [389] Yoneyama M, Fujita T: [Virus-induced expression of type I interferon genes]. *Uirusu* 2004, 54:161167.
- [390] Wimmer EA, Jle H, Pfeifle C, Cohen SM: A *Drosophila* homologue of human Sp1 is a head-specific segmentation gene. *Nature* 1993, 366:690694.
- [391] Estella C, Rieckhof G, Calleja M, Morata G: The role of buttonhead and Sp1 in the development of the ventral imaginal discs of *Drosophila*. *Dev. Camb. Engl.* 2003, 130:59295941.
- [392] Ing T, Tseng A, Sustar A, Schubiger G: Sp1 modifies leg-to-wing transdetermination in *Drosophila*. *Dev. Biol.* 2013, 373:290299.
- [393] Tone M, Tone Y, Babik JM, Lin C-Y, Waldmann H: The role of Sp1 and NF-kappa B in regulating CD40 gene expression. *J. Biol. Chem.* 2002, 277:88908897.

- [394] Zakrzewska A, Cui C, Stockhammer OW, Benard EL, Spaink HP, Meijer AH: Macrophage-specific gene functions in Spi1-directed innate immunity. *Blood* 2010, 116:e111.
- [395] Tone M, Powell MJ, Tone Y, Thompson SA, Waldmann H: IL-10 gene expression is controlled by the transcription factors Sp1 and Sp3. *J. Immunol. Baltim. Md 1950* 2000, 165:286291.
- [396] Larsson L, Thorbert-Mros S, Rymo L, Berglundh T: Interleukin-10 genotypes of the -1087 single nucleotide polymorphism influence sp1 expression in periodontitis lesions. *J. Periodontol.* 2011, 82:13761382.
- [397] Fan Y, Lee TV, Xu D, Chen Z, Lamblin A-F, Steller H, Bergmann A: Dual roles of Drosophila p53 in cell death and cell differentiation. *Cell Death Differ.* 2010, 17:912921.
- [398] Gowda PS, Zhou F, Chadwell LV, McEwen DG: p53 binding prevents phosphatase-mediated inactivation of diphosphorylated c-Jun N-terminal kinase. *J. Biol. Chem.* 2012, 287:1755417567.
- [399] Maqbool SB, Mehrotra S, Kolpakas A, Durden C, Zhang B, Zhong H, Calvi BR: Dampened activity of E2F1-DP and Myb-MuvB transcription factors in Drosophila endocycling cells. *J. Cell Sci.* 2010, 123:40954106.
- [400] Osta MA, Christophides GK, Vlachou D, Kafatos FC: Innate immunity in the malaria vector *Anopheles gambiae*: comparative and functional genomics. *J. Exp. Biol.* 2004, 207:25512563.
- [401] Kim LK, Choi UY, Cho HS, Lee JS, Lee W, Kim J, Jeong K, Shim J, Kim-Ha J, Kim Y-J: Down-Regulation of NF- $\kappa$ B Target Genes by the AP-1 and STAT Complex during the Innate Immune Response in Drosophila. *PLoS Biol.* 2007, 5:e238.

- [402] Moreno E, Basler K, Morata G: Cells compete for decapentaplegic survival factor to prevent apoptosis in *Drosophila* wing development. *Nature* 2002, 416:755759.
- [403] Shaw JP, Utz PJ, Durand DB, Toole JJ, Emmel EA, Crabtree GR: Identification of a putative regulator of early T cell activation genes. *Science* 1988, 241:202205.
- [404] Crzy F, Berta P, Girard F: Genome-wide analysis of Sox genes in *Drosophila melanogaster*. *Mech. Dev.* 2001, 109:371375.
- [405] Stathopoulos A, Van Drenth M, Erives A, Markstein M, Levine M: Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell* 2002, 111:687701.
- [406] Rovani MK, Brachmann CB, Ramsay G, Katzen AL: The dREAM/Myb-MuvB complex and Grim are key regulators of the programmed death of neural precursor cells at the *Drosophila* posterior wing margin. *Dev. Biol.* 2012, 372:88102.
- [407] Golling G, Li L, Pepling M, Stebbins M, Gergen JP: *Drosophila* homologs of the proto-oncogene product PEBP2/CBF beta regulate the DNA-binding properties of Runt. *Mol. Cell. Biol.* 1996, 16:932942.
- [408] Wei Q, Rong Y, Paterson BM: Stereotypic founder cell patterning and embryonic muscle formation in *Drosophila* require nautilus (MyoD) gene function. *Proc. Natl. Acad. Sci. U. S. A.* 2007, 104:54615466.
- [409] Fletcher JC, Burtis KC, Hogness DS, Thummel CS: The *Drosophila* E74 gene is required for metamorphosis and plays a role in the polytene chromosome puffing response to ecdysone. *Dev. Camb. Engl.* 1995, 121:14551465.
- [410] Lehmann M, Jiang C, Ip YT, Thummel CS: AP-1, but not NF-kappa B, is required for efficient steroid-triggered cell death in *Drosophila*. *Cell Death Differ.* 2002, 9:581590.

- [411] Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA: A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Res.* 2008, 36:25472560.
- [412] Berkeley *Drosophila* Genome Project (<http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl>).
- [413] Du Y.O, MSc Thesis: [http://uwspace.uwaterloo.ca/bitstream/10012/7586/1/Du\\_Yang\\_Olivia.p](http://uwspace.uwaterloo.ca/bitstream/10012/7586/1/Du_Yang_Olivia.p)
- [414] Nelson MR, Reisinger SJ, Henry SG: Designing databases to store biological information. *BIOSILICO* 2003, 1:134142.
- [415] Stein LD: Integrating biological databases. *Nat Rev Genet* 2003, 4:337345.
- [416] Yamashita R, Sugano S, Suzuki Y, Nakai K: DBTSS: DataBase of Transcriptional Start Sites progress report in 2012. *Nucleic Acids Res.* 2012, 40:D150-D154.
- [417] Gupta R, Bhattacharyya A, Agosto-Perez FJ, Wickramasinghe P, Davuluri RV: MPromDb update 2010: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-seq experimental data. *Nucleic Acids Res.* 2011, 39:D92D97.

# Chapter 6

## Appendices



## 6.1 Appendix one: calculation of N50 genome statistics for the *G.morsitans* draft assembly

With regards to genomics, N50 is a metric used to quantify the distribution of contig and scaffold/supercontig lengths within a draft assembly. Statistically, N50 can be described as the length N for which 50% of all bases in the sequences are in a sequence of length  $L < N$ . Thus N50 is a measure of the average length of a set of sequences, and the longer the N50 the better the assembly. The following are the R commands that were used to calculate the N50 for both contigs and scaffolds in the *G.morsitans* assembly.

**emboss commands for tab delimited infoseq file:**

```
infoseq -nocolumn -delimiter glossina-contigs-v1.fa >tsetse_contigs_infoseq
```

**R commands to read in Data and Check**

```
scaffold = read.table(file=file.choose(), header=T, sep=')
names(scaffold)
summary(scaffold)
summary(scaffold[6])
barplot(scaffold$Length)
```

**R commands for obtaining Nx Length**

```
analysis=rev(sort(scaffold$Length))
barplot(analysis)
n50 <-analysis[cumsum(analysis) >= sum(analysis)*0.5][1]
n50
```

**R commands to find the number of contigs / scaffold equal to N50 bp**

```
sum(analysis >n50)
```

### Function to calculate N10-N90

```
genome_stat = function(x,size)
y=rev(sort(scaffold$Length))
N_size=(size/100)
count_size=y[cumsum(y)>=sum(y)* N_size][1]
no_of_elements=sum(y>=count_size)
```

### Summary of results

(a) Scaffolds

Scaffold N length	Number of scaffolds	Length in Base pairs
N10	3	6177213
N20	35	594615
N30	137	266420
N40	312	172438
N50	570	120413
N60	928	89092
N70	1415	63149
N80	2126	41387
N90	3342	21132

(b) Contigs

Scaffold N length	Number of scaffolds	Length in Base pairs
N10	178	153105
N20	466	106972
N30	858	82070
N40	1363	62797
N50	2012	49769
N60	2838	38675
N70	3933	28468
N80	5479	19140
N90	8178	8767





## 6.2 Appendix two: graphical representations of quality scores before and after trimming

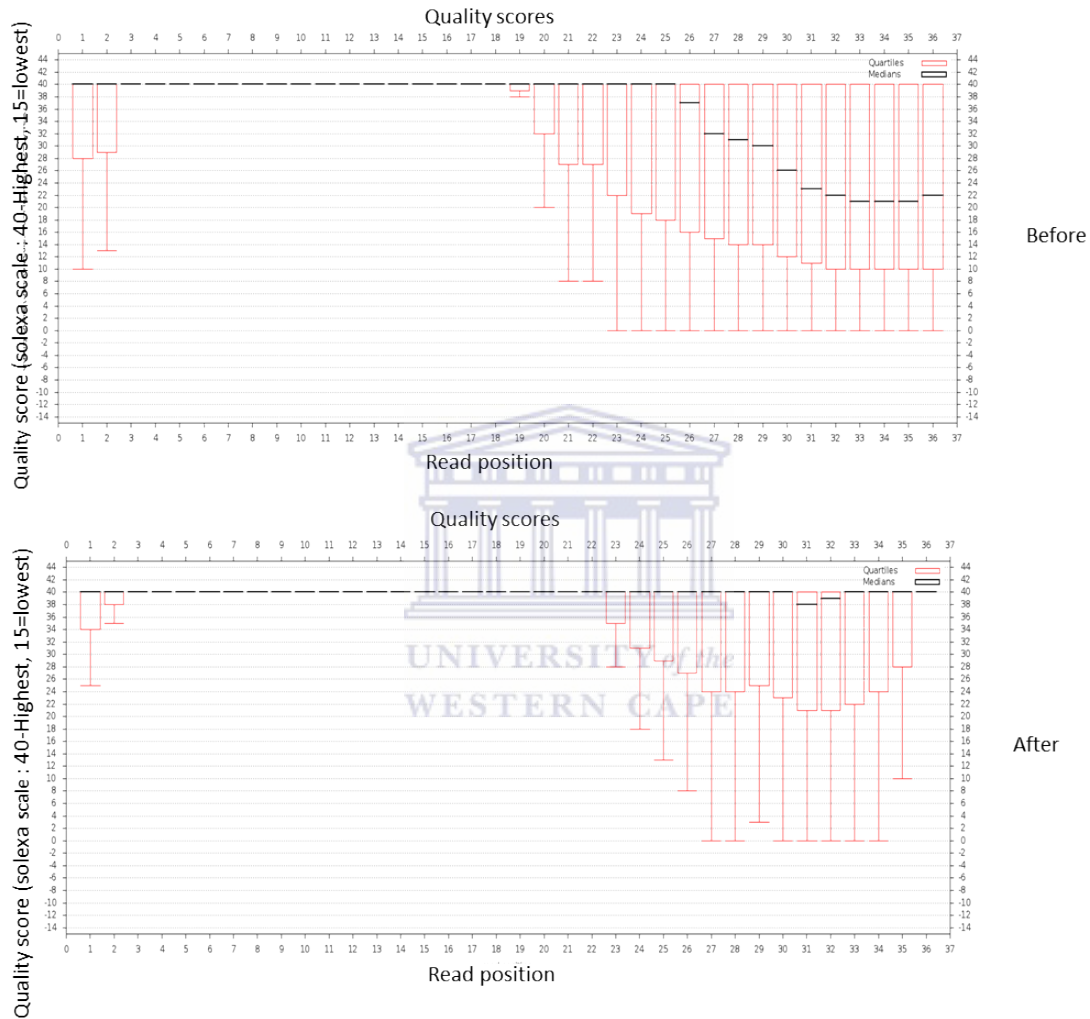


Figure 6.1: Quality chart of the TSS-seq reads before (i) and after (ii) quality control.

The x-axis indicates individual nucleotide position on the read whilst the y-axis depicts the quality scores. The whisker plots/quartiles (red) indicate the spread of the quality score across individual bases for the whole dataset whilst the black lined indicate the median of the quality score for each base pair across the dataset.

## 6.3 Appendix three: *G.morsitans* specific RNA POLII matrices and motif pictograms

### (i) TATA-Box

TATA-Box position-specific probability matrix.

A	C	G	T
0.407407	0.000000	0.444444	0.148148
0.111111	0.370370	0.296296	0.222222
0.000000	0.000000	0.518519	0.481481
0.000000	0.037037	0.074074	0.888889
0.518519	0.000000	0.000000	0.481481
0.000000	0.000000	0.000000	1.000000
0.666667	0.000000	0.000000	0.333333
0.259259	0.000000	0.000000	0.740741
0.962963	0.000000	0.000000	0.037037
0.370370	0.074074	0.000000	0.555556
0.592593	0.148148	0.222222	0.037037

TATA-Box pictogram:

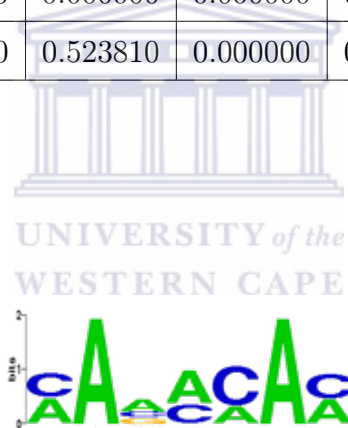


(ii) BRED

BRED position-specific probability matrix.

A	C	G	T
0.547619	0.119048	0.214286	0.119048
0.500000	0.500000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000
0.523810	0.238095	0.238095	0.000000
0.619048	0.380952	0.000000	0.000000
0.309524	0.690476	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000
0.476190	0.523810	0.000000	0.000000

BRED pictogram:

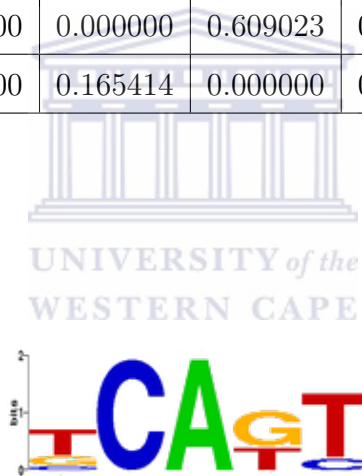


(iii) INR

INR position-specific probability matrix.

	A	C	G	T
1	0.293233	0.233083	0.150376	0.323308
2	0.300752	0.165414	0.165414	0.368421
3	0.278195	0.142857	0.240602	0.338346
4	0.000000	0.112782	0.255639	0.631579
5	0.000000	1.000000	0.000000	0.000000
6	1.000000	0.000000	0.000000	0.000000
7	0.000000	0.000000	0.609023	0.390977
8	0.000000	0.165414	0.000000	0.834586

INR pictogram:



(iv) MTE

MTE position-specific probability matrix.

A	C	G	T
0.000000	0.272727	0.000000	0.727273
0.136364	0.000000	0.000000	0.863636
0.000000	0.954545	0.045455	0.000000
0.409091	0.181818	0.090909	0.318182
0.363636	0.318182	0.181818	0.136364
0.545455	0.090909	0.000000	0.363636
0.136364	0.318182	0.090909	0.454545
0.090909	0.454545	0.454545	0.000000
0.909091	0.045455	0.045455	0.000000
0.681818	0.000000	0.318182	0.000000
0.590909	0.272727	0.090909	0.045455

MTE pictogram:



(v) DPE

DPE position-specific probability matrix.

	A	C	G	T
A	0.000000	0.500000	0.250000	0.250000
C	0.000000	0.000000	0.000000	1.000000
G	0.000000	0.000000	1.000000	0.000000
T	0.000000	0.250000	0.750000	0.000000
AA	0.000000	1.000000	0.000000	0.000000
CA	1.000000	0.000000	0.000000	0.000000
GA	1.000000	0.000000	0.000000	0.000000
TA	0.000000	0.500000	0.500000	0.000000

DPE pictogram:

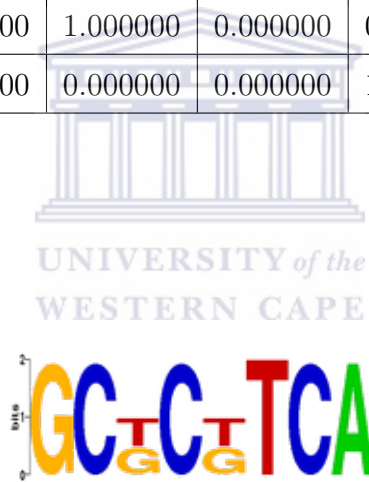


(vI) BREu

BREu position-specific probability matrix.

A	C	G	T
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	1.000000	0.000000
0.000000	1.000000	0.000000	0.000000
0.500000	0.000000	0.500000	0.000000
0.000000	1.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000

BREu pictogram:



## 6.4 Appendix four: Comparison of the number of immunity genes between *G. morsitans* and other dipterans

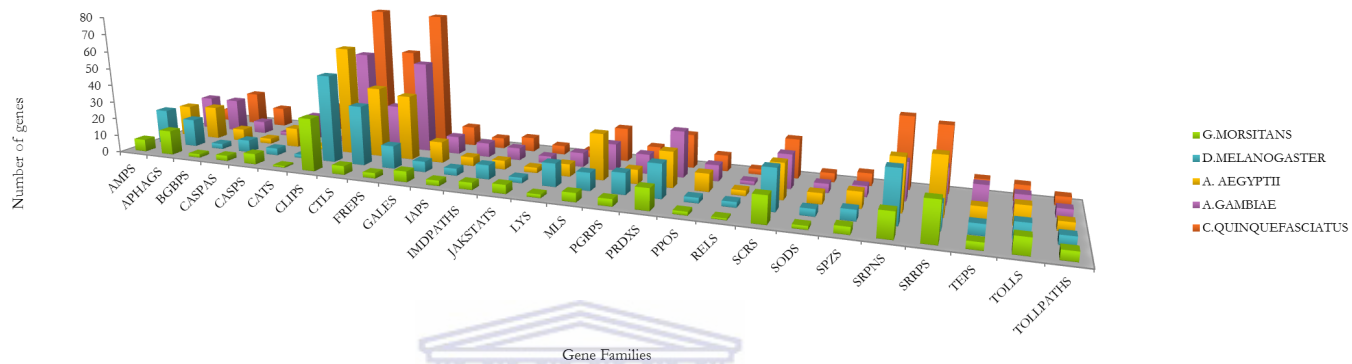


Figure 6.2: A graphical summary of the immunity gene families in selected dipterans.

