Table 5 : Tissue distribution of putative *Tsal* ESTs

PHRAP GENERATED CONTIG IDs	LENGTH (BP)	NUMBER OF READS								
		SG	MG	FB	PUM	PUF	TUF	*Tsal1	*Tsal2	TOTAL
1	1091	1								1
2	1161	1								1
3	1396	1								1
4	892	1								1
5	1198	1								1
6	699		1							1
7	1263								2	2
8	2197	2								2
9	886	2								3
10	2016	4								4
11	1202	4								4
12	1633	4	1							5
13	1463		3		2					5
14	1443	11								11
15	1436	13								13
16	1537	16								16
17	1396	17								17
18	1477	20								20
19	1516	22								22
20	1542	27								27
21	1576	32								32
22	1521	35								35
23	1584	55								55
24	1680	105	1							106
25	1755	115	1							116
26	2821	366	2							368
27	2715	452	1	1						454
28	2807	1011	4	1						1016
29	2495	1216	5	2	2	1	1	1		1228

The average contig length is 1599 base pairs. While the bulk of reads originate from the salivary glands, contig 29 has the widest tissue distribution.

SG – Salivary gland; PUM – Male pupae; MG – Midgut;

PUF – Female pupae; TUF – Combined tissue data; FB – Fat body;

**Tsal1* and *Tsal2* cDNA sequences were downloaded from GENBANK.

Table 6 : Conserved domains of *Tsal* ORFs

Contig 1D	Length (amino acids)	CDD	PFAM	SMART
1	117	Lac Z	Dihydropiridine Sensitive L type calcium channel	x
2	123	x	Shikimate quinate dehydrogenase	x
3	126	x	X	x
4	215	x	X	x
5	108	x	X	x
6	54	x	X	x
7	388	Nuclease	Endonuclease	Endonuclease
8	14	x	X	x
9	54	x	X	x
10	35	x	X	x
11	164	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
12	164	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
13	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
14	418	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
15	359	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
16	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
17	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
18	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
19	164	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
20	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
21	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
22	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
23	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
24	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
25	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
26	164	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
27	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
28	356	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
29	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease

Columns indicated by ‘x’ denote that no domains or motifs could be predicted with confidence.

UNIVERSITY of the
WESTERN CAPE

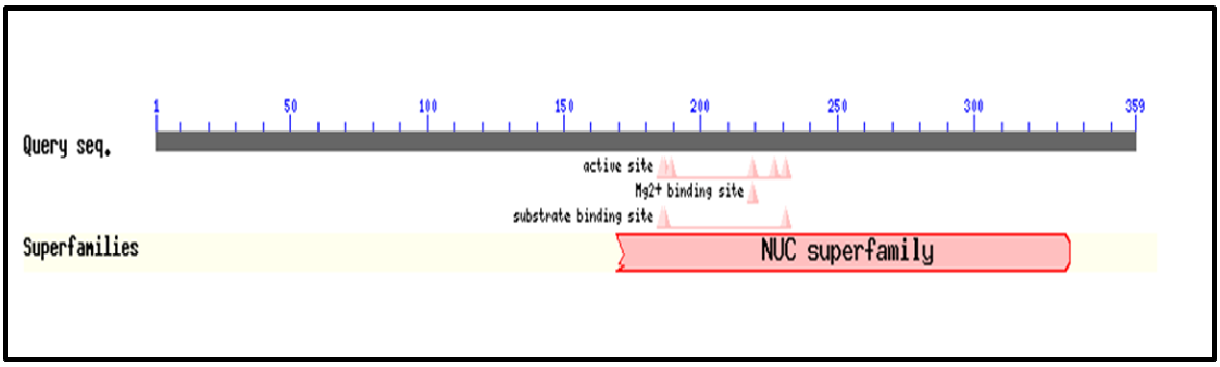


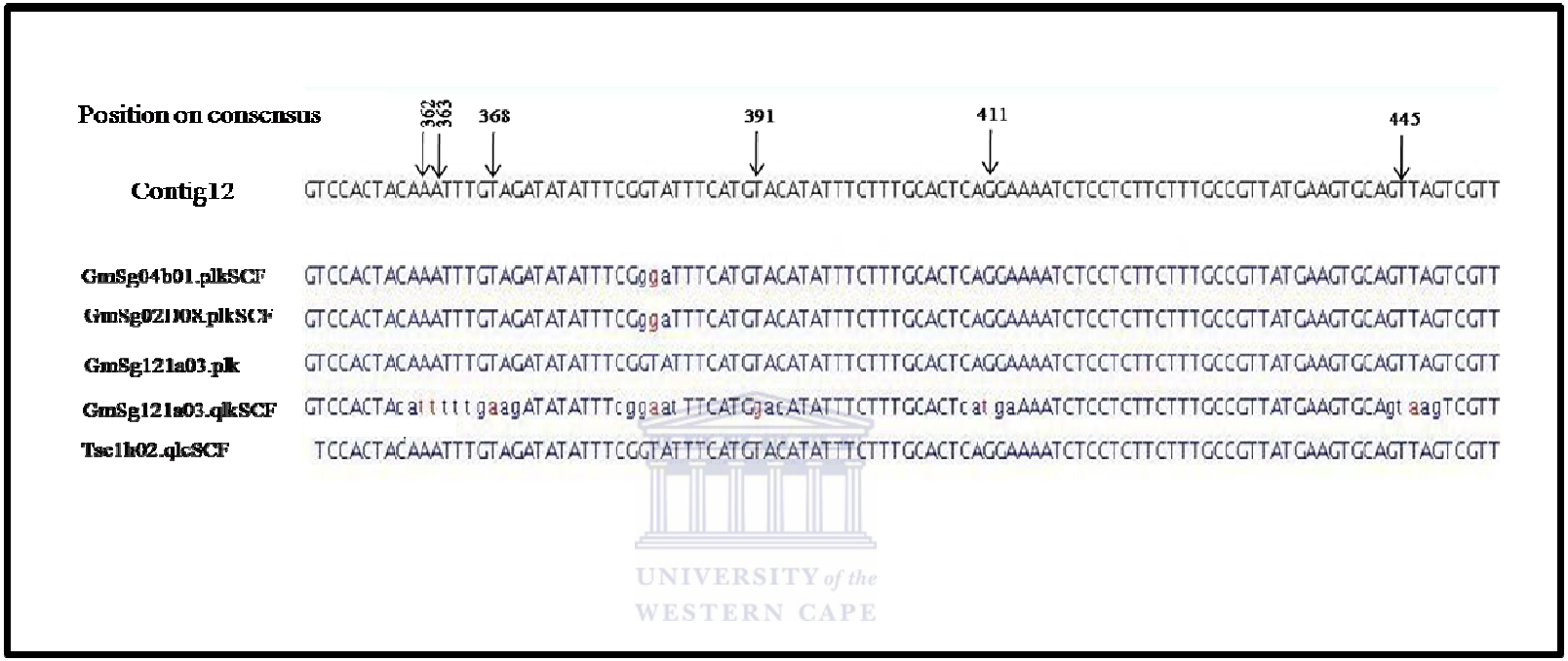
Figure 14: Graphical representation of the conserved domain for *G. morsitans* PHRAP generated contig29 putative *Tsal*.

Contig29 was generated as a result of PHRAP assembly of *Tsal* ESTs. A BLASTP search was carried out with contig29 ORF against NCBI non-redundant protein database. The grey bar corresponds to *G.morsitans* contig29 ORF. The nuclease superfamily conserved domain is depicted as the red bar. The relative positions of the substrate binding sites, Mg^{2+} binding site and the active site are indicated as peaks above the nuclease domain region.

3.9 Polymorphism in *Tsal* ESTs

UNIVERSITY of the
WESTERN CAPE

There was a high degree of variation observed within the reads in all PHRAP generated contigs. A predicted SNP is deemed valid only if the corresponding quality score on the base is satisfactory (Marth *et al.*, 1999). Additionally, a SNP is confirmed if its neighboring nucleotides are conserved over an interval. For example, Hayes and colleagues (2007) have proposed an interval of at least 50 bases. Single base mismatches were identified on contig12 using both a customized Perl script and the BEAP viewer (Figure 15). In addition, the corresponding quality scores of the putative SNPs were verified. Most of the variation observed was in the read GmSg121a03.plk with two instances of substitution of A to T at consensus positions 362 and 363. In addition, two instances of substitution from T to A occur at positions 368 and 445. Substitution for position 391 is T to G, while G is substituted with T in position 411(Table 7).



UNIVERSITY of the
WESTERN CAPE

Figure 15: Variation in the contig12 assembly.

This Figure represents consensus sequence of *Tsal* contig12 together with its constituent reads. The color differentiation option was selected to highlight putative SNPs. In this particular view putative SNPs are highlighted in red. Four reads denoted by the prefix "GmSg" originate from salivary gland are while one read (Tse1h02.qlcSCF) originates from the midgut. Most of the variation is observed in GmSg121a03.qlkSCF (see Table 7).

Table 7 : Putative SNPs in GmSg121a03.plk

Base position on consensus	Type of substitution	Quality score
362	A/T	12
363	A/T	15
368	T/A	10
391	T/G	10
411	G/T	15
445	T/A	12

UNIVERSITY of the
WESTERN CAPE

CHAPTER FOUR: DISCUSSION AND CONCLUSION

4.1 Discussion …… 57

4.1.1 Evolutionary relationships between insect serpins and *G.morsitans* putative serpins.. 57

4.1.2 Reactive center loops …… 60

4.1.3 Sequence variation in *Tsal* EST data …… 61

4.2 Conclusion and direction for future work …… 63

UNIVERSITY of the
WESTERN CAPE

4 DISCUSSION AND CONCLUSION

4.1 Discussion

4.1.1 Evolutionary relationships between insect serpins and *G.morsitans* putative serpins

Orthologous relationships established in this study are mainly with *Drosophila melanogaster* sequences. This is an observation that is shared with studies that have carried out functional annotation and/or phylogenetic analysis of sequences from insects such as *Anopheles gambiae, Manduca sexta, Bombyx mori* and *Glossina morsitans* (Christophides *et al.,* 2002; Zou *et al.,* 2009; Tanaka *et al.,* 2008; Lehane *et al.,* 2007). Both the genomes of *Anopheles gambiae* and *Drosophila melanogaster* have been sequenced and annotated (Holt et al., 2002; Celinker and Rubin 2003). It is possible that these genes have been deleted from the *Anopheles gambiae* genome during the evolutionary process.

A notable observation is the absence of a clear *G.morsitans* orthologue for *D.melanogaster serpin43Ac*, an extensively studied serpin. Despite extensive BLAST searches, a significant hit could not be obtained (data not shown). In their studies of the immune genes of *Anopheles gambiae* and *Bombyx mori*, Christophides *et al.*, (2002) and Tanaka *et al.,* (2008) did not identify an orthologue of *D.melanogaster serpin43Ac* for serpins from *Anopheles gambiae* and *Bombyx mori* respectively. *D.melanogaster serpin43Ac* (*nec*) has been implicated in the control of the toll mediated antifungal IMD pathway (Levashina *et al.,* 1999) by acting as an inhibitor of the clip domain serine protease, Persephone. The inference that can be drawn from this observation is that perhaps unlike *Drosophila melanogaster*, members of *Bombyx, Anopheline and Glossina* genera may not be overly challenged by fungal infections. Thus in their evolutionary processes, the functions of Persephone and *nec* may have diverged to serve other roles or cope with other regular challenges. Since genome sequencing is in progress, it is possible that the sequence of this gene is not yet available. Interestingly, *D.melanogaster serpin43Ac* shows weak sequence similarity with cn2215 (one of the sequences that was filtered with the presumption that it was a paralogue of cn2217). The putative orthologues of cn2217 are *A.gambiae serpin10* and *D.melanogaster serpin4*. In the event that it is established that cn2217 and cn2215 are indeed true paralogues and their functions are elucidated, the conclusion would be that the role of

D.melanogaster serpin43Ac diverged to perform the function of cn2215.

Different phylogenetic approaches confirm the clustering of cn2217 clusters with *A.gambiae serpin10* and *D.melanogaster serpin4*. Danielli and colleagues (2003) have established that alternative splicing of *A.gambiae serpin10* gives rise to four isoforms. In the same study, biochemical characterization of the inhibitory potential of three recombinant *serpin10* genes show that the isoforms are expressed in tissues involved in insect defense specifically, haemocytes and midgut epithelium. At least two isoforms are transcriptionally up-regulated during parasite establishment within the midgut potentially implicating them in antiparasitic action and/or parasite tolerance. *D.melanogaster serpin4* has been reported among the extensive arrays of genes that are upregulated in response to microbial challenge (Irving *et al.,* 2001). *D.melanogaster serpin4* gene encodes at least two different serpin proteins generated by alternative splicing of the last coding exon. This serpin was also shown to play a role in regulation of peptide maturation in *Drosophila* (Osterwalda *et al*., 2004). Interestingly, the spliced transcripts *Ag serpin10* and *Dm serpin4* segregate with contig 2217. It is likely that cn2217 also represents an alternative splice variant because of the other *Glossina* contigs (cn2215, cn2216, cn2218 and cn2220) that shared the same parent cluster with cn2217 (see section 2.3). Prediction and confirmation of alternative splice variants awaits the sequenced *Glossina* genome. It would be essential to establish the exact role that these homologues play in facilitating immune defenses in order to explain whether there has been pressure for diversification in order to deal with new challenges thus evolution by duplication followed by divergence, to produce a diverse set of paralogues.

Contig 6243 forms an orthologous group with two *Drosophila melanogaster* serpins deemed to be paralogues according to novel ENSEMBL predictions, *D.melanogaster serpin5* and *FBgn0038299(Dm38299)*. The topology of the clade containing cn6243 suggests that a second copy exists for *Glossina* or alternatively, that the second copy has been deleted over time. These paralogues would have arisen before tsetse flies and *Drosophila* species diverged. Besides facilitating wing expansion during early *Drosophila* development (Charron *et al*., 2008), *D.melanogaster serpin5* has been shown to be upregulated in response to immune stimulation by oligonucleotide DNA microarrays (Irving *et al.,* 2001). Within the same cluster is *A.gambiae*

serpin9. Currently there is no data on the exact physiological roles of genes in this cluster namely, *D.melanogaster serpin5, FBgn0038299 (Dm38299)* and *A.gambiae serpin9*. Homology modeling of the 3D alignment of α-1- antitrypsin and cn6243 depicts conservation of elements of secondary structure including residues at the reactive center loop. Comparative modeling is important in establishing evolutionary relationships as it enables the detection of sequence similarity between the target and the template. The 3D structure of proteins from the same family is more conserved than the amino acid sequences. Presence of most of the elements of secondary structure including the reactive center loop, show that indeed cn6243 may be a complete representative of the cognate gene. Thus, this serpin presents an excellent candidate for biochemical and genetic experiments in an effort to understand its exact role.

G.morsitans cn1692 forms an orthologous group with *D.melanogaster serpin27A* and *A.gambiae serpin2*. *D.melanogaster serpin27A* and *A.gambiae serpin2* have been implicated in regulation of the melanization cascade (Michel *et al.*, 2005) by facilitating the site-specific localization of the ensuing melanotic response during the insect's immune response. The target specificity of cn1692 is probably the same as that of *D.melanogaster serpin27A* as they both possess identical scissile bond residues (see Figure 13).

UNIVERSITY of the
WESTERN CAPE

A search of the serpin domain in cn3298 revealed that this domain is represented by two fragments (see Figure 7). Domain-(ii) exhibits extensive coiling thus it is predicted to be a transmembrane portion. A BLAST search against *Drosophila melanogaster* proteins on ENSEMBL identified FBgn0031973 as the best hit. Interestingly, cn3298 gives the highest percent identity against its corresponding insect serpin queries whereas the percentage coverage is poor (see Table 3). Thus, a high percent identity is not always an indicator of sufficient sequence coverage due to divergence across a gene with localized conserved regions. The fragmentation of cn3298 domain into two portions may account for poor amino acids overlap in the alignment. The amino acid sequences flanking the reactive center loop of cn3298 and selected serpins (see Figure 13) are conserved, indicating that it may indeed be an inhibitory serpin. The congruence of the phylogenetic tree reconstruction is perturbed by cn3298 to some degree, a fact that could be attributed to poor amino acid overlap in the alignments.

4.1.2 Reactive center loops

The sequence and conformation of the RCL largely determines the selectivity of inhibition, thus sequence alignments of the serpin C terminus comprising the RCL are particularly revealing (see Figure 13). The RCL has received much attention during the study of serpin mechanism of action for the reason that it holds the critical amino acids considered to control much but not all of the inhibitory specificity of the serpins. In this study, we were able to establish a high degree of conservation in the amino acid environment flanking the reactive center loop, indicating that these serpins may indeed be inhibitory. In addition, we were able to determine the position of the scissile bond (P'_1 P_1). The scissile bond residues are identical for cn1692 and its orthologue *D.melanogaster serpin27A.* Therefore, cn1692 may have a regulatory role in *G. morsitans* defense responses, similar to *D.melanogaster serpin27A* because of common target specificity. We could not establish the P_1 residue of the scissile bond for the cn6243. Albeit the high degree of conservation in the region spanning the RCL the scissile bonds of cn2217 and cn3298 are not identical, a strong indication that they may not have the same target specificity. However, this cannot be used to rule out the possibility that these contigs are indeed inhibitory serpins. In addition, this observation can be attributed to the evolutionary process of these genes that may have been accompanied by mutations that led to changes in the amino acid composition. The high degree of conservation of the RCL is an indicator of evolutionary conservation of serpin genes that participate in the signal transduction pathways. Genes in the signal transduction pathways have remained highly conserved throughout the different insect orders (Tanaka *et al.*, 2008). Conserved gene evolution of insect signal transduction repertoires reflects the essential requirement of these genes for common immune mechanisms against infectious microorganisms.

UNIVERSITY of the
WESTERN CAPE

4.1.3 Sequence variation in *Tsal* EST data

Tissue distribution of ESTs shows that *Tsal* transcripts are not only expressed in the salivary gland but in other tissues, for example, the fat body and midgut (Table 5). While Li and colleagues (2001) had established that *Tsal1* expression is limited to adult and larval stages, the results of the assembly show that *Tsal1* assembles with three transcripts that originated from the pupae. The assembly of *Tsal1* with the highest number of transcripts with a varied tissue distribution indicates the possibility of multiple isoforms for *Tsal* that may not have been characterized.

Homology searches against NCBI non-redundant protein database identified *Anopheles gambiae* and *Culex quinquefasciatus* endonucleases as the next most significant hits albeit low percentage identity. This is consistent with the study performed by Calvo and Ribeiro (2006) on the sialotranscriptome of *Culex quinquefasciatus* whereby *Tsal1* and *Tsal2* exhibit similarities with an abundant cluster of sequences coding for secreted proteins with endonuclease activity. Sixty nine percent of the contigs generated by PHRAP assembly of putative *Tsal* ESTs are predicted to be nucleases more specifically endonucleases, by SMART. PFAM denotes them to be either DNA or RNA endonucleases. However, contigs in which less than five ESTs assembled (excluding contig 7) are predicted to encode other conserved domains, for example, Dihydropiridine and Shikimate quinate dehydrogenase (see table 6). The underlying sequences in these contigs do not share a high level of similarity with the rest of the ESTs in the cluster thus their assembly as singletons and possible possession of variant conserved domains.

UNIVERSITY of the
WESTERN CAPE

We observe a high degree of variation in the contigs. The high degree of variation coupled with low quality scores for contig twelve's single base pair mismatches hampered the identification of specific SNPs. Accurate identification of *Tsal* SNPs requires resequencing of salivary gland ESTs to obtain accurate base calls. In addition, *Tsa1l*, *Tsal2a* and *Tsal2b* do not assemble as one contig indicating that their underlying sequences of potential alternative splice variants do not have a high degree of similarity. Variation in the *Tsal* cluster is concordant with Lanzaro and colleagues (1999) who have proposed that salivary gland proteins of blood-sucking dipterans may have unusual degree of polymorphism because the insect may benefit from antigenic

variation of its salivary proteins. Our findings could be used in advancing the understanding of tsetse salivary gland protein evolution.

In their study, Ghosh and Mukopadhyay (1998) demonstrate that natural or host anti-sandfly salivary gland antigen immunization reduces sand fly's blood feeding efficiency and increases mortality in the post bloodmeal vector population. They attribute this to the fact that anti-sandfly saliva immune sera probably bind with the respective antigen-presenting sites of the sandfly salivary gland components causing the sandfly death. A similar study on tsetse flies by Caljon *et al.*, (2006) demonstrates that, immunization with recombinant *Tsal1* (*rTsal1*) does not lead to increased tsetse mortality. The implication of this is that perhaps the *Tsal* protein does not possess antigen binding sites.

The majority (69%) of all conceptually translated putative *Tsal* ORFs are predicted to encode the nuclease domain thus it is our interest to discover the role of nucleases in organisms specifically in dipterans. In *Drosophila melanogaster*, DNase II deficiency has been shown to impair innate immunity function by reducing total haemocyte numbers (Seong *et al.*, 2006). In humans, DNase II was shown to have a role in immunity protection where genetic information may be transferred from apoptotic cells to living cells after being phagocytosed, a potential threat to the genome (Bergsmedh *et al.*, 2006). We propose that *Tsal* may have an immune role (indeed there are transcripts from the fat body which is the major immune response organ) in the tsetse fly and that probably its deposition causes degradation of the apoptotic cells of the human host cell to prevent integration of host DNA into the tsetse fly. This may be an acquired function in the evolutionary process of the tsetse fly to protect itself and it may be linked with the vector competence of the tsetse flies. Apoptosis is usually accompanied by chromosomal DNA degradation and is important for homeostasis of metazoans (Liu *et al.*, 2008). The existence of multiple isoforms may be a mechanism to deal with a wide array of hosts for the tsetse fly.

4.2 Conclusion and direction for future work

With members of the serpin superfamily present over all major taxonomic groups, structural diversification of serpins has occurred to fulfill their roles in myriad of biological processes. Thus, several studies have focused on serpin evolution in all the taxonomic groups. In this study, we focused on a comparative genomics approach to identify putative serpins orthologues from unpublished data of the *Glossina morsitans* transcriptome. By use of data mining and phylogenetic reconstruction, we have gained insight into possible biological functions of putative serpins. However, all these data remains to be confirmed by genetic and biochemical experiments, a process that will be successfully carried out once we have a reference genome.

Ideally, one would want to look at how the immune repertoires of the parasites and *Glossina morsitans* have coevolved to establish whether the evolutionary process has facilitated vector refractoriness. For example, mapping of the contigs onto chromosomes and subsequent identification of gene architectures to provide an insight into the intron-exon organization as well as relative positions of the gene regulatory machinery. This data provides an impetus into structural and functional studies of this family of proteins with a specific focus on the putative paralogues. Paralogues are very important in studying the evolutionary process because they derive from gene family expansions and gene losses, or cases of exceptionally high sequence divergence where they acquire other functions. In the event that these paralogues have evolved to perform different functions then multiple regulatory elements present in the promoter region will need to be dissected.

UNIVERSITY of the
WESTERN CAPE

Studies on the evolutionary patterns of the immune repertoire of *Drosophila melanogaster* and disease vector mosquitoes has established that signal modulation uses a vast reservoir of serine proteases and their inhibitors. Most of these genes have evolved species-specific functions with gene duplication and expansions among certain genes. Complete *Glossina morsitans* serpin sequences would provide insights in the context of a wider array of members of insect serpins in terms of their evolutionary patterns. In addition complete gene architectures will help identify whether these serpins harbor signal peptides either at the N or C terminus to give insight into cellular localization of the proteins.

Identification of the exact function of *Tsal*, for example, by RNA*i* will be particularly informative because it is unique to *Glossina* thus it may offer some insight into the exclusive pathway of *Glossina* evolution. Probably, *Tsal* genes may have diversified to gain other functions that may not be present in other haematophagous insects. This may be an evolutionary process that helps *Glossina morsitans* carve its niche.

UNIVERSITY of the
WESTERN CAPE

REFERENCES

Aksoy S., Berriman M., Hall N., Hattori M., Hide W., Lehane M.J., (2005). A case for a *Glossina* genome project. *Trends Parasitology*. **21(3):**107-11.

Aksoy S., Maudlin I, Dale C., Robinson A.S., O'Neill S.L., (2001). Prospects for control of African trypanosomiasis by tsetse vector manipulation. *Trends in Parasitology.* **17(1):**29-35.

Allsopp R. (2001). Options for vector control against trypanosomiasis in Africa. *Trends in Parasitology*. **17:**15-19.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990). "Basic local alignment search tool". *J Mol Biol*. **215(3):**403–410.

Bergsmedh A., Szeles A., Henriksson M., Bratt A., Folkman M.J., Spetz A.L., Holmgren L . (2001). Horizontal transfer of oncogenes by uptake of apoptotic bodies. *Proc Natl Acad Sci U S A*. **98(11):**6407-11.

Berriman M., Ghedin E., Hertz-Fowler C., Blandin G., Renauld H., Bartholomeu D.C., Lennard N.J., Caler E., Hamlin N.E., Haas B *et al.* (2005). The genome of the African trypanosome *Trypanosoma brucei*. *Science*. **309 (5733):** 416-22.

Bonaldo M.F., Lennon G., Soares M.B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* **6(9):**791-806.

Boulanger N., Brun R., Ehret-Sabatier L., Kunz C., Bulet P. (2002). Immunopeptides in the defense reactions of *Glossina morsitans* to bacterial and *Trypanosoma brucei brucei* infections. *Insect Biochem Mol Biol*. **32(4):**369-75.

UNIVERSITY of the
WESTERN CAPE

Burke J., Davison D., Hide W. (1999). d2_cluster: a validated method for clustering EST and full-length cDNA sequences. *Genome Res.* **9(11):**1135-42.

Caljon G., Van Den Abbeele J., Sternberg J.M., Coosemans M., De Baetselier P., Magez S. (2006). Tsetse fly saliva biases the immune response to Th2 and induces anti-vector antibodies that are a useful tool for exposure assessment. *Int J Parasitology.* **36(9):**1025-35.

Calvo E., Ribeiro J.M. (2006). A novel secreted endonuclease from Culex quinquefasciatus salivary glands. *J Exp Biol.* **209:**2651-9.

Cecchi G., Mattioli R.C., Slingenbergh J., de la Rocque S. (2008). Land cover and tsetse fly distributions in sub-Saharan Africa. *Med Vet Entomol*. **22(4):**364-73.

Celniker S.E., Rubin G.M. (2003). The *Drosophila melanogaster* genome. *Annu Rev Genomics Hum Genet.* **4:**89-117.

Charron Y., Madani R., Combepine C., Gajdosik V., Hwu Y., Margaritondo G'., Vassalli J.D.(2008). The serpin Spn5 is essential for wing expansion in *Drosophila melanogaster*. *Int J Dev Biol* **7:** 933- 942.

Cheng Q., Aksoy S., (1999). Tissue tropism, transmission and expression of foreign genes in vivo in midgut symbionts of tsetse flies. *Insect Molecular Biology.* **8(1):**125-132.

Christoffels A., van Gelder A., Greyling G., Miller R., Hide T., Hide W. (2001). STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res.* **29(1):**234-8.

Christophides G.K., Zdobnov E., Barillas-Mury C., Birney E, Blandin S., Blass C., Brey P.T., Collins F.H., Danielli A., Dimopoulos G.*et al.* (2002). Immunity-related genes and gene families in *Anopheles gambiae*. *Science.* **298:**159-165.

Clamp M., Cuff J., Searle S.M., Barton G.J., (2004). The Jalview Java alignment editor. *Bioinformatics*. **20(3):**426-7.

Clark M.D., Hennig S., Herwig R., Clifton S.W., Marra M.A., Lehrach H., Johnson S.L. (2001). An Oligonucleotide Fingerprint Normalized and Expressed Sequence Tag Characterized Zebrafish cDNA Library *Genome Res*. **11:**1594-1602.

Clarke E.P., Cates G.A., Ball E.H., Sanwal B.D., (1991). A collagen-binding protein in the endoplasmic reticulum of myoblasts exhibits relationship with serine protease inhibitors. *J Biol Chem*. **266(26):**17230-5.

Cooper D.N., Smith B.A., Cooke H.J., Niemann S., Schmidt J. (1985). An estimate of unique DNA sequence heterozygosity in the human genome. *Hum Genet.* **69(3):**201-5.

Danielli A., Kafatos F.C., Loukeris T.G. (2003). Cloning and characterization of four *Anopheles* gambiae serpin isoforms, differentially induced in the midgut by Plasmodium berghei invasion. *J Biol Chem*. **278(6):**4184-93.

De Gregorio E., Han S.J., Lee W.J., Baek M.J., Osaki T., Kawabata S., Lee B.L., Iwanaga S., Lemaitre B., Brey P.T. (2002). An immune-responsive Serpin regulates the melanization cascade in *Drosophila. Dev Cell*. **4:**581-92.

DeLano W.L. (2002). The PyMOL Molecular Graphics System on the World Wide Web (http://www.pymol.org).

Dumas M and Bouteille B. (1996). Human African trypanosomiasis. *C R Seances Soc Biol Fil*. **190 (4):**395-408.

El-Sayed N.M., Hegde P., Quackenbush J., Melville S.E., Donelson J.E. (2000). The African trypanosome genome. *Int J Parasitol*. **30(4):**329-45.

UNIVERSITY of the
WESTERN CAPE

Ellis J.A., Shapiro S.Z., ole Moi-Yoi O., Moloo S.K., (1986). Lesions and saliva-specific antibody responses in rabbits with immediate and delayed hypersensitivity reactions to the bites of *Glossina morsitans centralis. Vet Pathol.* **23(6):**661-7.

Epidemiology and control of African trypanosomiasis. Report of a WHO committee. Geneva Switzerland: World Health Organization Technical report series, 1986:739.

Ewing B., Green P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome.* **8(3):**186-94.

Ewing B., Hillier L., Wendl M.C., Green P. (1998). Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. *Genome Res*. **8(3):**175-85.

Fairlamb A.H. (2003). Chemotherapy of human African trypanosomiasis: current and future prospects. *Trends in Parasitology*. **19(11):** 488-494.

Felsenstein J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*. **5:** 164-166.

Filipa F., Jorge C., Andreia F., Nicolas N., Pedro N., Agustin B., João P., Maria O., Jorge S., Jorge A., Sónia C. (2008). An alternative approach to detect Trypanosoma in *Glossina* (Diptera, *Glossinidae)* without dissection. *J Infect Developing Countries.* **2(1):**63-67.

Gascuel O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol.* **14:**685–695.

Geiger A., Ravel S., Frutos R., Cuny G. (2005). Sodalis glossinidius (Enterobacteriaceae) and vectorial competence of *Glossina palpalis gambiensis* and *Glossina morsitans morsitans* for *Trypanosoma congolense* savannah type. *Curr Microbiol*. **51(1):**35-40.

UNIVERSITY of the
WESTERN CAPE

Ghosh K.N, Mukhopadhyay J. (1998). The effect of anti-sandfly saliva antibodies on *Phlebotomus argentipes* and *Leishmania donovani*. *Int J Parasitol.* **28(2):**275-81.

Green P. (1999). PHRAP Documentation http://www.phrap.org/

Guindon S., Gascuel O. (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." *Systematic Biology*. **52(5):**696-704.

Guryev V., Berezikov E., Malik R., Plasterk R.H., Cuppen E. (2004) Single nucleotide polymorphisms associated with rat expressed sequences. *Genome Res*. **14(7):**1438-43.

Haines L.R, Hancock R.E., Pearson T.W. (2003). Cationic antimicrobial peptide killing of African trypanosomes and Sodalis glossinidius, a bacterial symbiont of the insect vector of sleeping sickness. *Vector Borne Zoonotic Dis*. **3(4):**175-86.

Hajduk S.L. (1984). Antigenic variation during the developmental cycle of *Trypanosoma brucei*. *J Protozool*. **31(1):**41-7.

UNIVERSITY of the
WESTERN CAPE

Hammond G.L., Smith C.L., Goping I.S., Underhill D.A., Harley M.J., Reventos J., Musto N.A., Gunsalus G.L., Bardin C.W. (1987) Primary structure of human corticosteroid binding globulin, deduced from hepatic and pulmonary cDNAs, exhibits homology with serine protease inhibitors. *Proc Natl Acad Sci U S A.* **84(15):**5153-7.

Hao Z., Kasumba I., Lehane M.J., Gibson W.C., Kwon J., Aksoy S. (2001). Tsetse immune responses and trypanosome transmission: implications for the development of tsetse-based strategies to reduce trypanosomiasis. *Proceedings of the National Academy of Sciences of the United States of America.* **98(22):**12648–12653.

Hayes B.J., Nilsen K., Berg P.R., Grindflek E., Lien S. (2007). SNP detection exploiting multiple sources of redundancy in large EST collections improves validation rates. *Bioinformatics*. **23(13):**1692-3.

Herwig R., Schulz B., Weisshaar B., Hennig S., Steinfath M., Drungowski M., Stahl D., Wruck W., Menze A., O'Brien J., Lehrach H., Radelof U. (2002). Construction of a 'unigene' cDNA clone set by oligonucleotide fingerprinting allows access to 25 000 potential sugar beet genes. *Plant J*. **32:**845–857.

Hide G. (1999). History of sleeping sickness in East Africa. *Clin Microbiol Rev.* **12(1):**112-25.

Hoffman J.A. (2003). The immune response of *drosophila*. *Nature* **426:** 33-38.

Holt R.., Subramanian G.M., Halpern A., Sutton G.G., Charlab R., Nusskern D.R., Wincker P., Clark A.G., Ribeiro J.M., Wides R *et al.*(2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science.* **298(5591):**129-49.

Hopkins P.C., Carrel R.W., Stone S.R. (1993). Effects of mutations in the hinge region of serpins. *Biochemistry*. **32:**7650-7657.

Hu C. and Aksoy S. (2006). Innate immune responses regulate trypanosome parasite infection of the tsetse fly *Glossina morsitans morsitans*. *Mol Microbiol*. **60(5):**1194-204.

Huang X and Madan A. (1999). CAP3:A DNA sequence assembly program. *Genome Res*. **9(9):**868-77.

Huelsenbeck, J. P. and Ronquist F. (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics*. **17:**754-755.

Irving J.A., Pike R.N., Lesk A.M., Whisstock J.C. (2000). Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. *Genome Res*. **12:**1845-1864.

Irving P., Troxler L., Heuer T.S., Belvin M., Kopczynski C., Reichhart J.M., Hoffmann J.A., Hetru C. (2001). A genome-wide analysis of immune responses in *Drosophila. Proc Natl Acad Sci*. **98(26):**15119–15124.

Jones D.T., Taylor W.R., Thornton J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences (CABIOS).* **8(3):**275-282.

Kabayo J.P. (2002). Aiming to eliminate tsetse from Africa. *Trends in Parasitology.* **18:**473-475.

Kasahara M., Naruse K., Sasaki S., Nakatani Y., Qu W., Ahsan B., Yamada T., Nagayasu Y., Doi K., Kasai Y., Jindo T, *et al.* (2007). The Medaka draft genome and insights into vertebrate genome evolution. *Nature*. **447(7145):** 714-9.

Kennedy P.G.E. (2007). The fatal sleep. Endiburgh, United Kingdom: Luath Press limited.

Koltes J.E., Hu Z.L., Fritz E., Reecy J.M., (2009). BEAP: The BLAST Extension and Alignment Program- a tool for contig construction and analysis of preliminary genome sequence. *BMC Res Notes*. **22(2):**11.

Kumar S., Tamura K., Nei M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.* **5(2):**150-63.

Lanzaro G.C., Lopes A.H., Ribeiro J.M., Shoemaker C.B., Warburg A., Soares M., Titus R.G. (1999). Variation in the salivary peptide, maxadilan, from species in the Lutzomyia longipalpis complex. *Insect Mol Biol*. **8(2):**267-75.

Lehane M.J., Gibson W., Lehane S.M. (2007). Differential expression of fat body genes in *Glossina morsitans morsitans* following infection with *Trypanosoma brucei brucei. Int J Parasitol.* **38(1):**93-101.

UNIVERSITY of the
WESTERN CAPE

Levashina E.A., Langley E., Green C., Gubb D., Ashburner M., Hoffmann J.A., Reichhart J.M. (1999). Constitutive activation of toll-mediated antifungal defense in serpin-deficient *Drosophila. Science.* **285:**1917-1919.

Li S., Kwon J., Aksoy S. (2001). Characterization of genes expressed in the salivary glands of the tsetse fly *Glossina morsitans morsitans*. *Insect Molecular biology* **10(1):**69-76.

Lindblad-Toh K., Wade C.M., Mikkelsen T.S., Karlsson E.K., Jaffe D.B., Kamal M., Clamp M., Chang J.L., Kulbokas E.J 3rd, Zody M.C., *et al.* (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature.* **438(7069):**803-19.

Liu M.F., Wu X.P., Wang X.L., Yu Y.L., Wang W.F., Chen Q.J., Boireau P., Liu M.Y. (2008). The functions of Deoxyribonuclease II in immunity and development. *DNA Cell*. **27(5):**223-8.

Löbermann H., Tokuoka R., Deisenhofer J., Huber R. (1984). "Human a1-proteinase inhibitor. Crystal structure analysis of two crystal modifications, molecular model and preliminary analysis of the implications for function." *J. Mol. Biol.* **177:** 531-556.

Mant M.J and Parker K.R. (1981) Two platelet aggregation inhibitors in tsetse (*Glossina*) saliva with studies of roles of thrombin and citrate in in-vitro platelet aggregation. *Br J Haematol*. **48(4):**601-8.

Marth G.T., Korf I., Yandell M.D., Yeh R.T., Gu Z., Zakeri H., Stitziel N.O., Hillier L., Kwok P.Y., Gish W.R. (1999). A general approach to single-nucleotide polymorphism discovery. *Nat Genet.* **23(4):**452-6.

Michel K., Budd A., Pinto S., Gibson T.J., Kafatos F.C. (2005). *Anopheles gambiae* SRPN2 facilitates midgut invasion by the malaria parasite Plasmodium berghei. *EMBO Rep*. **6(9):**891-7.

Miller R.T., Christoffels A.G., Gopalakrishnan C., Burke J., Ptitsyn A.A., Broveak T.R., Hide W.A. (1999). A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.* **9(11):**1143-55.

Milligan P.J., Maudlin I., Welburn S.C. (1995). Trypanozoon: infectivity to humans is linked to reduced transmissibility in tsetse II. Genetic mechanisms. *Exp. Parasitol.* **81:**409–415.

Nagaraj S.H., Gasser R.B., Ranganathan S. (2007). A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics.* **8(1):**6-21.

Notredame C., Higgins D., Heringa J (2000). T-Coffee: A novel method for multiple sequence alignments. *Journal of Molecular Biology*. **302:** 205-217.

Okolo C.J., Jenni L., Molyneux D.H., Wallbanks K.R. (1990). Surface carbohydrate differences of *Glossina* salivary glands and infectivity of *Trypanosoma brucei gambiense* to *Glossina*. *Ann Soc Belg Med Trop*.**70(1):**39-47.

Osterwalder T., Kuhnen A., Leiserson W.M., Kim Y.S., Keshishian H. (2004). *Drosophila* serpin 4 functions as a neuroserpin-like inhibitor of subtilisin-like proprotein convertases. *J Neurosci.* **24(24):**5482-91.

Pays E., Lips S., Nolan D., Vanhamme L., Pérez-Morga D. (2001). The VSG expression sites of *Trypanosoma brucei*: multipurpose tools for the adaptation of the parasite to mammalian hosts. *Mol Biochem Parasitol*.**114(1):**1-16.

Pearson W.R. (1997). Identifying distantly related protein sequences. *Comput Appl Biosci.* **13:** 325-332.

Pinder M., Fumoux F., van Melick A., Roelants G.E (1987). The role of antibody in natural resistance to African trypanosomiasis. *Vet Immunol Immunopathol*. **17(1-4):**325-32.

UNIVERSITY of the
WESTERN CAPE

Potempa J., Korzus E., Travis J. (1994). The Serpin Superfamily of Proteinase Inhibitors: Structure, Function, and Regulation. *J. Biol. Chem*. **269:** 15957-15960.

Rawlings N.D., Tolle D.P., Barrett A.J. (2004). Evolutionary families of peptidase inhibitors. *Biochem J*. **378:**705–716.

Ribeiro J.M., Charlab R., Pham V.M., Garfield M., Valenzuela J.G. (2004). An insight into the salivary transcriptome and proteome of the adult female mosquito *Culex pipiens quinquefasciatus. Insect Biochem Mol Biol*. **34(6):**543-63.

Rickman L.R. (1977). Variation in the test responses of clone derived *Trypanosoma* (Trypanozoon) *brucei brucei* and T.(T).b. *rhodesiense* relapse antigenic variants, examined by a modified blood incubation infectivity test and its possible significance in Rhodesian sleeping sickness transmission. *Med J Zambia*.**11(2):**31-7.

Rudd S. (2003). Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci*. **8(7):**321-9.

UNIVERSITY of the
WESTERN CAPE

Saitou N. and Nei M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4(4):406-25.

Sali A and Blundell T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol .234(3):779-815.

Seong C.S., Varela-Ramirez A., Aguilera R.J. (2006). DNase II deficiency impairs innate immune function in *Drosophila*. *Cell Immunol*. **240(1):**5-13.

Silverman G.A., Bird P.I., Carrell R.W., Church F.C., Coughlin P.B., Gettins P.G., Irving J.A., Lomas D.A., Luke CJ., Moyer R.W., Pemberton P.A., Remold-O'Donnell E., Salvesen G.S., Travis J., Whisstock J.C. (2001). The Serpins Are an Expanding Superfamily of Structurally Similar but Functionally Diverse Proteins. *J Biol. Chem.* **276:** 33293-33296.

Staden R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6(7):**2601-10.

Stich A., Abel P.M., Krishna S. (2002). Human African trypanosomiasis. *BMJ.* **27; 325(7357):**203-6.

Stickney H.L., Schmutz J., Woods I.G., Holtzer C.C., Dickson M.C., Kelly P.D., Myers R.M., Talbot W.S.(2002). Rapid mapping of zebrafish mutations with SNPs and oligonucleotide microarrays. *Genome Res.***12(12):**1929-34.

Tanaka H., Ishibashi J., Fujita K., Nakajima Y., Sagisaka A., Tomimoto K., Suzuki N., Yoshiyama M., Kaneko Y., Iwasaki T., Sunagawa T., Yamaji K., Asaoka A., Mita K., Yamakawa M. (2008). A genome-wide analysis of genes and gene families involved in innate immunity of *Bombyx mori. Insect Biochem Mol Biol.* **38(12):**1087-110.

Tanji T., Hu X., Weber AN., Ip YT. (2007). Toll and IMD Pathways Synergistically Activate an Innate Immune Response in *Drosophila melanogaster*. *Molecular and Cellular Biology.* **27(12):**4578-4588.

Tanji T and Ip Y.T. (2005). Regulators of the Toll and Imd pathways in the *Drosophila* innate immune response. *Trends in immunology.* **26(4):** 193-198.

Thompson J.D., Higgins D.G., Gibson T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22bghj(22):** 4673-80.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A *et al.* (2001). The sequence of the human genome. *Science.* **29(5507):** 1304-1351.

UNIVERSITY of the
WESTERN CAPE

Vickerman K. (1985). Developmental cycles and biology of pathogenic trypanosomes. *Br Med Bull*. **41(2):** 105-14.

Waterhouse A.M., Procter J.B., Martin D.M., Clamp M., Barton G.J. (2009). Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics. [Epub ahead of print].

Welburn S.C., Maudlin I., Ellis D.S. (1998) Rate of trypanosome killing by lectins in midguts of different species and strains of *Glossina. Med Vet Entomol*. **3(1):**77-82.

Welburn S.C., Maudlin I. (1999). Tsetse-Trypanosome interactions: Rites of passage. *Parasitology today*. **15(10):**399-403.

Whisstock J.C., Pike R.N., Jin L., Skinner R., Pei X.Y., Carrell R.W., Lesk A.M (2000) Conformational changes in serpins: II. the mechanism of activation of antithrombin by heparin. *Journal of Molecular Biology*. **301(5):**1287-1305.

Zou Z., Picheng Z., Weng H., Mita K., Jiang H. (2009). A comparative analysis of serpin genes in the silkworm genome. *Genomics.* **93(4):**367-75.

UNIVERSITY of the
WESTERN CAPE

APPENDICES

APPENDIX I: Outline of PHRED and PHRAP

Conversion of chromatograms to computer readable form by PHRED

Conversion of DNA sequencer trace files to computer readable form is the first step of any assembly process; this was done by the program PHRED (Ewing and green, 1998). PHRED reads the DNA sequencer trace data and examines the peaks around each base to assign a quality score. The scores provide a measure of sequence quality and are calculated using the logarithmic function obtained from Ewing and Green (1998).

$$q=10 * \log 10 (P)$$

Where;

q = quality score

p = probability that a particular base was an error

Each base call is therefore logarithmically linked to error probability. Quality scores range from 4 to 60 with higher values corresponding to better sequence quality. For example, a quality score of 30 represents 1/1000 chances of being incorrect while a score of 20 represents 1/100 chances of being incorrect. PHRED scores can therefore be used to extract portions of high quality in any given sequence or the entire sequence depending with the biological question being answered during downstream EST analysis.

Due to the inherent nature of errors that are associated with any EST sequencing project bases within the middle segment of a given EST sequence typically contain the high quality bases while the extreme ends normally contain vector contaminants and repeats. PHRED then writes the base calls in FASTA, PHD (text files that consist of base call and quality information, used during contig editing) or SCF (Standard Chromatogram File) formats. The quality scores are written to FASTA or PHD files (Ewing and Green, 1998; Ewing *et al.*, 1998).

UNIVERSITY of the
WESTERN CAPE

Summary of the steps in a PHRAP assembly

1. PHRAP reads the sequence from the input file as well as the corresponding quality file.

2. PHRAP identifies and trims homopolymeric regions (those that consist almost entirely a single base). These regions are most likely due to poor data quality and may give rise to false matches.

3. PHRAP identifies all potentially overlapping pairs of sequences. Two sequences must have at least one "word" (whose default is set at 14 bases) and a minimum alignment score (default is set at 30).

4. PHRAP computes adjusted quality scores for each base in each overlapping read taking into consideration read orientation and sequencing chemistry. In addition PHRAP calculates log-likelihood ratio (LLR) scores. LLR scores compare the hypothesis that the reads truly overlap to the hypothesis that they are from 95% similar repeats. A pairwise match tends to have a positive LLR score if the two reads overlap, whereas it tends to have a negative LLR score if the two reads are from different repeats. Possible problem clones like chimeric and deletion reads are also identified and withheld from the assembly.

UNIVERSITY of the
WESTERN CAPE

5. PHRAP merges reads into contigs (consensus) starting with pairwise overlaps with the highest LLR scores. The consensus sequence is "pieced" together from individual sequence reads with the highest adjusted quality score at any base.

APPENDIX II: List of Perl scripts

Perl script for creating quality scores for three *Tsal* cDNAs downloaded from GENBANK.

```
#!/usr/bin/perl
use warnings  ;
use strict ;

# Takes a fasta file, determines the length of the sequence
# and creates bogus phred quality scores for each sequence
# The default phred score is 15
# Usage: perl <scriptname> fastafile1 fastafile2....

foreach my $file(@ARGV){
  my $qual_out = $file.".qual" ;
  open(FASTA_IN, $file) ;
  open(QUAL_OUT, ">>$qual_out") ;
  $/ = "\n>" ;          # This line will make the 'while' loop below
deal
                # with a complete FASTA entry at a time
  while(my $sequence = <FASTA_IN>){
    $sequence =~ s/\>//g ;        # remove all ">" characters
    my @array = split(/\n/, $sequence) ;
    my $head = shift(@array) ;
    my $sequence_string = join("", @array) ;
    print              QUAL_OUT             phred_score_producer($head,
length($sequence_string)), "\n" ;
  }
  close(FASTA_IN) ;
  close(QUAL_OUT) ;
}

# SUBROUTINES
sub phred_score_producer{
  my ($header, $sequence_length) = @_ ;
```

```
  my $fasta_entry = ">$header" ;
  for( my $index = 0 ; $index < $sequence_length ; $index++ ){
    if(($index/20) == int($index/20)){           # this makes sure my
lines
                                          # are 20 characters long
      $fasta_entry .= "\n30" ;
    }else{ $fasta_entry .= " 30" ;
    }
  }
  return($fasta_entry) ;
}
```

UNIVERSITY of the
WESTERN CAPE

Perl script for extracting read names from a ".ace" file output

```
#! /usr/bin/perl

$/="\nCO";# stop at a line that ends with CO
my $file = shift @ARGV;# initialises our file as an array
open(FH $file);

while (<>) {
	next if ($_=~/^AS/);#match stdin with anything that starts with AS
	my @lines = split("\n", $_);#create an array that will contain the lines
	my $contigname = shift @lines;#read the first element of the lines array and store it in $contigname
	$contigname =~/^\s+(\S+)/;#in the store for contig name pick up the second item (S+) after the first space s+
	my $contig = $1; #take the pattern in the first set of brackets which is also equal to $1 and put it into my new variable called $contig.
	foreach my $line(@lines) {
		if ($line =~/^AF\s+(\S+)/) {
			my $readname = $1;
			print "$contig\t$readname\n";
		}
	}
}
```

UNIVERSITY of the
WESTERN CAPE

Perl script for extracting single base pair mismatches from a ".GDE" file.

```
#! /usr/bin/perl
#
# background: test.pl reads in an ace file and generates a list of contigIDs vs ESTnames. eg., Contig1 Est1 est2 est3
# --output file= Ace_ESTnames
#
# acefile converted to GDE format using AMOS. However the estnames have a number and brackets at the end (ESTname(447))
# --output file =tsal.fasta.gde
#
# stripped off the number at the end of the name and saved the sequences into a new file using ConvertGdeName.pl
# /ConvertGdeName.pl tsal.fasta.gde3 ----output was tsal.fasta.gde3.new
#
# Use bioperl index script to index the output of ConvertGdeName.pl -index_file called GDESEQ
# usage: ~/cvs_src/bioperl-live/scripts/index/bp_index.PLS -dir . GDESEQ ./tsal.fasta.gde3.new
#
# SNP_count.pl
# reads in output from test.pl and store the information in a hash table.
#
# found bug - est can go beyond contiglength
use strict;
use Bio::SeqIO;
use Bio::Seq;
use Bio::Index::Fasta;

my $estnames = shift;
my %hash;
```

UNIVERSITY of the
WESTERN CAPE

```
my              $in              =              Bio::Index::Fasta->new(-
filename=>"/home/sarah/tsal/SNP_results/TSALSEQ");
#my $out = Bio::SeqIO->new(-file=>">$f.test", -format=>"Fasta");

open(F, $estnames);
while (<F>) {
      chomp;
      my @i = split(/\s+/);
      my $contig = shift @i;
      @{$hash{$contig}} = @i;
}
close(F);

foreach my $contig(keys %hash) {
      #print "$contig\n";
      my $contig_seq = $in->fetch($contig);
      #print $contig_seq->id."\n";

      my $contiglen = length($contig_seq->seq);
      my @est_comp;
      foreach my $est(@{$hash{$contig}}) {
            my $est_seq = $in->fetch($est);
            my ($start_string, $end_string);
            if ($est_seq->desc =~/\<(.*)\>$/) {
                  my $str = $1;
                  #print "$str\n";
                  # start and stop are pos where EST aligns with other
ESTs
                  # these are positions relative to the consensus
                  my ($start, $stop) = split(/\s+/, $str);
                  #print "start = $start stop=$stop\n";
                  if ($start > 1) {
                        my $start_string_len = ($start - 1);
                        #print "$start_string_len\n";
```

UNIVERSITY of
WESTERN CAPE

```
                $start_string = "";
                foreach my $j(1..$start_string_len) {
                        $start_string .="N";
                }
        } else {
                $start_string = "";
        }

        #deal with the end_string
        if ($stop < $contiglen) {
                my $end_string_len = ($contiglen - $stop);
                $end_string = "";
                foreach my $j(1..$end_string_len) {
                        $end_string .="N";
                }
        } else {
                $end_string="";
        }
    }
    #print "$est .$start_string.\n";
    #print "$est .$end_string\n";
    my $new_est = $start_string.$est_seq->seq().$end_string;
    #print "$est     .$new_est.\n";

    my @new_estseq = split("", "\U$new_est");
    push @est_comp, [@new_estseq];
}
my ($j, $i, %temphash);
for ($j=0; $j< $contiglen; $j++) {
%temphash=();
for ($i=0; $i<= $#est_comp; $i++) {
#print $dna[$i][$j]."\n";
#$temphash{$est_comp[$i][$j]}++ if ($est_comp[$i][$j] ne "N");
$temphash{$est_comp[$i][$j]}++;
```

```
                #print "$i:$j .$est_comp[$i][$j]. - ";
        }
        if (scalar(keys %temphash) ==1) {
             #conserved pos
             print "pos-idx-$j";
                  my ($n) = keys %temphash;
                  print " $n ***\n";
        } else {
             print "pos-idx-$j";
           foreach my $base(keys %temphash) {
             print " $base\[$temphash{$base}\]";
           }
           print "\n";
        }
       }
}
```

UNIVERSITY of the
WESTERN CAPE