

**Reptilian Beta-defensins**

Paul Mason-Smith

Thesis presented for the degree of  
Doctor of Philosophy  
Birmingham City University

OCTOBER 2023

## **Declaration**

I declare that the research and analysis presented in this thesis is my own unless otherwise stated. This work has not been submitted for any other degree and this thesis was composed entirely by me.

Paul Mason-Smith

October 2023

## Abstract

Beta-defensins are a family of small cationic antimicrobial peptides (AMPs) that are strongly conserved throughout Eukaryotes. Their discovery was associated with broad spectrum antimicrobial activity, but it has since been determined that this family of proteins have more diverse functions, including roles such as a chemokine for macrophages and determining coat colour in dogs.

In Reptiles, the innate immune response is the principal system for clearing invading pathogenic organisms. Beta-defensins are a major component of the innate immune response, yet their biological function is largely unknown. It is likely that their antimicrobial activity mimics that of their mammalian counterparts. Due to little research within reptilian beta-defensins many questions remain to be answered. Cluster characterisation, gene polymorphisms and defensin gene repertoire along with evolutionary mechanisms, conservation of synteny and physical properties of these peptides needs to be elucidated.

This work identified novel gene clusters in 3 snake, 3 lizard, and 2 turtle/tortoise genomes as well as additional analysis on 2 Crocodylia species. Bioinformatic discovery through techniques such BLAST and annotation analysis reveals that these genes reside in single cluster. Conservation of synteny is observed within these clusters and in all species the cluster was flanked on one side by Cathepsin B (CTSB) and the other side by either, Exportin 1 (XPO1) in the Squamates and by Translocation Associated Membrane Protein 2 (TRAM2) in Crocodylia and the turtles/tortoises. Throughout reptilians, gene homology is observed with genes residing nearest CTSB being most obvious and through gene annotation and splice site prediction a two-exon organisation is the predominant gene structure of these beta-defensins. Their physical properties suggest that they are mostly cationic, with a few detected exceptions. The  $dS/dN$  ratio of synonymous and nonsynonymous substitutions on a gene level suggests that there is conservation in the first exon, which encodes a signal peptide and more positive Darwinian selection in the second exon which encodes the active AMP suggesting an 'arms race' between host and pathogen. Site wise evolutionary analyses on the second exon using HyPhy shows a common beta-defensin cysteine motif which is undergoing purifying selection and residues between these, showing positive selection. All

this demonstrates that these are an evolving group of immune genes. In addition to this, a streamlined, cost-effective methodology has been developed to express, purify, and study these peptides to aid the investigation into their physical properties. This can provide information as to whether these novel antimicrobials have a much-needed use in future therapeutics or medicines.



## Acknowledgments

Firstly, I would like to thank Dr David Hughes for giving me this wonderful opportunity. Secondly, I would like to thank Dr David Lee for his advice during my studies and Dr Loukia Tsaprouni for taking the time to critically read through this work. Finally, a massive thanks to Dr Irmgard Hausmann for the huge amount of support you have given me over the last few years of this study. I wouldn't have made it this far if you hadn't given your time to help me and get me over the finish line.

## Table of Contents

Declaration	ii
Abstract	iii
Acknowledgements	v
Table of Contents	vi
List of figures	xi
List of Tables	xvii
Abbreviations	xix

### **Chapter 1: Introduction**

<b>1.1</b> A need for discovery	1
<b>1.2.1</b> Initial discovery and Characterisation	1
<b>1.2.2</b> Structure and characteristics	2
<b>1.2.3</b> Thermal and pH stability	5
<b>1.3</b> Evolutionary origins of beta-defensins	6
<b>1.3.1</b> Gene cluster dynamics and formation	6
<b>1.3.2</b> Selection pressures	9
<b>1.4</b> Role of beta-defensins in innate immunity	10
<b>1.4.1</b> Toll-like Receptors – Membrane bound sensors	11
<b>1.4.2</b> NOD-like Receptors – Intracellular Cytoplasmic sensors	11
<b>1.4.3</b> RIG-like receptors – Intracellular viral sensors	12
<b>1.4.4</b> Beta-defensins and PRR recognition and activation	15
<b>1.4.5</b> Role in wound healing	17
<b>1.5</b> Beta-defensin Discovery and experimental research methods	19
<b>1.5.1</b> Data mining approach	19
<b>1.5.2</b> Sequences alignment and Phylogeny analysis	21
<b>1.5.3</b> Protein expression and purification	22
<b>1.5.4</b> Antimicrobial Activity	24
<b>1.5.5</b> Immobilisation of AMPs on surface coatings	28
<b>1.5.6</b> Potential and current uses of AMPs in medical and food industries	29
<b>1.6</b> Reptiles – An interesting, untapped resource	30

### **Chapter 2: Method Development for Data Mining and Gene Annotations of Genomic DNA to Establish Beta-defensin Clusters**

<b>2.1</b> Initial Searches	33
<b>2.2</b> First Annotations	37
<b>2.3</b> Gene finding programs for Identification	40
2.3.1 GENSCAN	40
2.3.2 FGENESH – Softberry, Inc.	41
<b>2.4</b> Splice Site Prediction	45
<b>2.5</b> Repeat Masking	49
<b>2.6</b> Additional changes during development of methodology pipeline	52
<b>2.7</b> Evaluation and comparison of search strategy from Santana, F. L. et al. (2021)	53

## Chapter 3: Crocodylia

<b>3 Aims</b>	57
<b>3.1 <i>Crocodylus porosus</i></b>	57
3.1.1 Data Mining and Cluster Assembly	58
3.1.2 Cluster Organisation and Beta-defensin Sequences	59
3.1.3 Physical properties	60
3.1.4 Selection analysis	61
3.1.5 Repeat Sequence Landscape	62
<b>3.2 <i>Alligator mississippiensis</i></b>	66
3.2.1 Data Mining and Cluster Assembly	67
3.2.2 Cluster organisation and Beta-defensin sequences	68
3.2.3 Physical properties	69
3.2.4 Selection analysis	70
3.2.5 Repeat Sequence Landscape	73
<b>3.3 Conservation of Synteny</b>	73
<b>3.4 Summary</b>	76

## Chapter 4: Lizards

<b>4. Aims</b>	77
<b>4.1 <i>Podarcis muralis</i></b>	77
4.1.1 Data Mining and Cluster Assembly	78
4.1.2 Cluster organisation and Beta-defensin sequences	79
4.1.3 Physical properties	80
4.1.4 Selection analysis	83
4.1.5 Repeat Sequence Landscape	85
<b>4.2 <i>Lacerta agilis</i></b>	87
4.2.1 Data Mining and Cluster Assembly	87
4.2.2 Cluster organisation and Beta-defensin sequences	88
4.2.3 Physical properties	89
4.2.4 Selection analysis	90
4.2.5 Repeat Sequence Landscape	94
<b>4.3 <i>Zootoca vivipara</i></b>	96
4.3.1 Data Mining and Cluster Assembly	96
4.3.2 Cluster organisation and Beta-defensin sequences	97
4.3.3 Physical properties	98
4.3.4 Selection analysis	99
4.3.5 Repeat Sequence Landscape	102
<b>4.4 Conservation of Synteny</b>	107
<b>4.5 Summary</b>	107

## Chapter 5: Snakes

<b>5. Aims</b>	108
<b>5.1 <i>Crotalus viridis viridis</i></b>	108
5.1.1 Data Mining and Cluster Assembly	109
5.1.2 Cluster organisation and Beta-defensin sequences	110
5.1.3 Physical properties	111
5.1.4 Selection analysis	112
5.1.5 Repeat Sequence Landscape	113
<b>5.2 <i>Naja naja</i></b>	116
5.2.1 Data Mining and Cluster Assembly	116
5.2.2 Cluster organisation and Beta-defensin sequences	118
5.2.3 Selection analysis	120
5.2.4 Physical properties	122
5.2.5 Repeat Sequence Landscape	123
<b>5.3 <i>Thamnophis elegans</i></b>	124
5.3.1 Data Mining and Cluster Assembly	125
5.3.2 Cluster organisation and Beta-defensin sequences	126
5.3.3 Physical properties	130
5.3.4 Selection analysis	131
5.3.5 Repeat Sequence Landscape	132
<b>5.4 Conservation of Synteny</b>	134
<b>5.5 Summary</b>	140

## Chapter 6: Turtle/Tortoise

<b>6. Aims</b>	141
<b>6.1 <i>Chelonia Mydas</i></b>	141
6.1.1 Data Mining and Cluster Assembly	142
6.1.2 Cluster organisation and Beta-defensin sequences	143
6.1.3 Physical properties	144
6.1.4 Selection analysis	145
6.1.5 Repeat Sequence Landscape	148
<b>6.2 <i>Gopherus evgoodei</i></b>	150
6.2.1 Data Mining and Cluster Assembly	150
6.2.2 Cluster organisation and Beta-defensin sequences	152
6.2.3 Physical properties	152
6.2.4 Selection analysis	154
6.2.5 Repeat Sequence Landscape	154
<b>6.3 Conservation of Synteny</b>	160
<b>6.4 Summary</b>	160

## Chapter 7: Python Beta-defensin 1 (MSBD1) Experimental

7.1 DNA extraction and RF cloning of MSBD1	161
7.2 Electroporation and transformation.	164
7.3 Sequencing of plasmid insert and isolated MSBD1 gene.	165
7.4 Expression and solubility testing	168
7.5 MPB affinity chromatography purification	170
7.6 Tobacco Etch Virus Protease (TEVp) cleavage of fusion MBP-MSBD1	173
7.7 Purification with Immobilised Metal Affinity Chromatography (IMAC)	175
7.8 Refolding and Purification of MSBD1 using Size Exclusion Chromatography	177
7.9 Summary	180

## Chapter 8: Discussion 183

### Critical Reflection: 196

## Chapter 9: Material and Methods

9.1 Computer based resources	199
9.1.1 Multiple Sequence Alignments	199
9.1.2 Phylogenetic trees	199
9.1.3 Signal peptide prediction	199
9.1.4 Selection analysis	199
9.1.5 Gene-Wise Selection analysis	200
9.1.6 Site-wise Selection Analysis	200
9.1.7 Genome DNA Alignment Software	200
9.1.8 DNA sequencing data	201
9.1.9 Protein Characterisation	201
9.2 LABORATORY PROCEDURES	201
9.2.1 DNA Extraction Protocol for Shed Reptile Skins	201
9.2.2 Plasmid and expression strain selection	202
9.2.3 Restriction Free Cloning	202
9.2.4 Primers	205
9.2.5 PCR Reaction Protocols	205
9.2.6 Agarose Gel Electrophoresis	206
9.2.7 Protocol for preparation of electrocompetent <i>E. coli</i> for transformation of plasmid in to holding strain DH5 $\alpha$	206
9.2.8 Recombinant plasmid purification from holding strain for transformation into SHuffle <sup>®</sup> Competent <i>E. coli</i>	207
9.2.9 Plasmid Preparation into SHuffle <sup>®</sup> Competent <i>E. coli</i> .	207
9.2.10 DNA sequencing of transformed <i>E. coli</i>	208
9.2.11 Expression of Python Beta-defensin in SHuffle <sup>®</sup> <i>E. coli</i> .	209
9.2.12 Cell lysis for Beta-defensin recovery	209
9.2.13 Solubility Testing for expressed MSBD1	209
9.1.14 Affinity Purification of fusion MPB-PBBD1 fusion protein	210
9.1.15 SDS-PAGE analysis	210

9.1.16 Staining Procedure	211
9.1.17 Cleavage of MBP fusion protein to release PBBD1	212
9.1.18 Immobilised metal ion affinity (IMAC) chromatography for purification of PBBD1.	212
9.1.19 Refolding and purification by Size Exclusion Chromatography	213
9.1.20 Preliminary Antimicrobial Activity testing	213
9.1.21 Diafiltration	213
<b>Chapter 10: References</b>	<b>214</b>
<b>Chapter 11: Appendices</b>	<b>233</b>

## List of Figures

<b>Figure 1.1</b> Beta-defensin genomic organisation.	3
<b>Figure 1.2</b> Schematic illustration of AMP interaction with lipid membranes.	4
<b>Figure 1.3</b> Origin of new gene copies through gene duplication.	8
<b>Figure 1.4</b> Putative mechanism for defensin gene duplication.	9
<b>Figure 1.6</b> TLRs, RLHs, and NLRs in renal cells.	14
<b>Figure 1.7</b> Models for the interaction of Beta-defensins with cell surface receptors.	15
<b>Figure 1.8.</b> Principle functions of defensins in inflammatory and defensive reactions against pathogens.	19
<b>Figure 2.1:</b> Chicken Beta-defensin concatemer used as template for initial searches.	34
<b>Figure 2.2.</b> BLAST search results from single Beta-defensin used in query sequence.	35
<b>Figure 2.3.</b> BLAST search result from using Chicken concatemer.	36
<b>Figure 2.4.</b> Page showing how region on scaffold is changed.	37
<b>Figure 2.5.</b> Results from BLAST search from smaller 850kbp stretch of Scaffold.	38
<b>Figure 2.6.</b> Input box for FATSA file sequence in EMBOSS Sixpack program website.	39
<b>Figure 2.7.</b> Positions of potential first and second exon match.	39
<b>Figure 2.8.</b> Output file from EMBOSS Sixpack.	40
<b>Figure 2.9.</b> Genomic Map of BLAST matches.	41
<b>Figure 2.10</b> Input field for FGENESH.	43
<b>Figure 2.11.</b> Identification of Potential Beta-defensin gene missing from GenScan analysis.	43

<b>Figure 2.12.</b> Full genomic map of Predicted Genes.	45
<b>Figure 2.13</b> Splice site prediction input box with DNA sequence.	46
<b>Figure 2.14.</b> Splice site prediction output file.	46
<b>Figure 2.15.</b> Donor Splice site annotated on EMBOSS file.	47
<b>Figure 2.16.</b> Acceptor splice site annotated on EMBOSS output.	47
<b>Figure 2.17.</b> Poly adenylation signal highlighted on EMBOSS file.	49
<b>Figure 2.18.</b> Summary of repeats identified from repeat masker.	50
<b>Figure 2.19.</b> Output file showing positions of repeats on the scaffold of interest.	50
<b>Figure 2.20.</b> Positions of first Beta-defensin prediction.	51
<b>Figure 2.21</b> Input of Beta-defensin in excel spreadsheet.	52
<b>Figure 2.22</b> Gap exceeding 3000bp in cluster region between whole Beta-defensin genes.	52
<b>Figure 3.1</b> Genomic organisation of the <i>C. porosus</i> Beta-defensin cluster.	57
<b>Figure 3.2</b> Multiple sequence alignment of <i>C. porosus</i> beta-defensins cluster.	60
<b>Figure 3.3</b> Amino acid sequence logo of second exon.	60
<b>Figure 3.4.</b> Ratio of synonymous and nonsynonymous substitutions.	61
<b>Figure 3.5</b> Genomic organisation of the <i>A. mississippiensis</i> Beta-defensin cluster.	65
<b>Figure 3.6</b> Ratio of synonymous and nonsynonymous substitutions in <i>A. mississippiensis</i> .	68
<b>Figure 3.7</b> Multiple sequence alignment of <i>A. mississippiensis</i> beta-defensin sequences.	71
<b>Figure 3.8</b> Amino acid sequence Logo of the second exon in <i>A. mississippiensis</i> .	71
<b>Figure 3.9</b> Conservation of Synteny.	72



<b>Figure 4.1</b> Genomic organisation of the <i>P. muralis</i> Beta-defensin cluster.	79
<b>Figure 4.2</b> Dot plot of the cluster region and genomic organisation of <i>P. muralis</i> genes.	80
<b>Figure 4.3</b> Multiple sequence alignment of 80 Beta-defensin genes identified in the <i>P. muralis</i> cluster.	82
<b>Figure 4.4</b> Ratio of synonymous and nonsynonymous substitutions in <i>P. muralis</i> .	83
<b>Figure 4.5</b> Amino acid sequence logo of second exon of <i>P. muralis</i> .	84
<b>Figure 4.6</b> Phylogenetic tree of the DNA coding sequences for <i>P. muralis</i> .	85
<b>Figure 4.7</b> Genomic organisation of the <i>L. agilis</i> Beta-defensin cluster.	89
<b>Figure 4.8</b> Proportions of synonymous and nonsynonymous substitutions in <i>L. agilis</i> .	91
<b>Figure 4.9</b> Multiple sequence alignment of 64 Beta-defensin genes identified in the <i>L. agilis</i> cluster.	92
<b>Figure 4.10</b> Amino acid sequence Logo of the second exon peptide of <i>L. agilis</i> .	93
<b>Figure 4.11</b> Phylogenetic tree of the DNA coding sequences of <i>L. agilis</i> .	94
<b>Figure 4.12</b> Genomic organisation of the <i>Z. vivipara</i> Beta-defensin cluster.	98
<b>Figure 4.13</b> Multiple sequence alignment of Beta-defensin genes identified in the <i>Z. vivipara</i> cluster.	100
<b>Figure 4.14</b> Ratio of synonymous and nonsynonymous substitutions in <i>Z. vivipara</i> .	101
<b>Figure 4.15</b> Amino acid sequence logo of <i>Z. vivipara</i> second exons.	102
<b>Figure 4.16</b> Dot plots of Lizard species.	104
<b>Figure 4.17</b> Synteny between Lizard clusters.	104
<b>Figure 4.18</b> Cluster alignment of each cluster region in Lizards.	105
<b>Figure 4.19</b> Cluster alignment of each cluster region produced by sequences matches produced by MAUVE sequence aligner.	106
<b>Figure 5.1</b> Multiple sequence alignment of <i>Crotalus viridis viridis</i> beta-defensins.	112

<b>Figure 5.2</b> Genomic organisation of the Beta-defensin cluster of <i>Crotalus viridis viridis</i> .	112
<b>Figure 5.3</b> Proportion of synonymous and nonsynonymous substitutions in <i>C. v. viridis</i> .	113
<b>Figure 5.4</b> Amino acid sequence logo of mature peptides of <i>C. v. viridis</i> .	114
<b>Figure 5.5</b> Genomic organisation of the <i>Naja naja</i> beta-defensin cluster.	118
<b>Figure 5.6</b> Multiple sequence alignment of <i>Naja naja</i> beta-defensins.	119
<b>Figure 5.7</b> Phylogenetic tree of the DNA coding sequences of exons 1 and 2 of <i>Naja naja</i> .	120
<b>Figure 5.8</b> Proportion of synonymous and nonsynonymous substitutions in <i>Naja naja</i> .	121
<b>Figure 5.9</b> Amino acid sequence logo of mature peptides in <i>Naja naja</i> .	122
<b>Figure 5.9</b> Genomic organisation of the <i>Thamnophis elegans</i> beta-defensin cluster	127
<b>Figure 5.10</b> Dot plot of the cluster region of <i>T. elegans</i> .	128
<b>Figure 5.11</b> Multiple sequence alignment of <i>Thamnophis Elegans</i> beta-defensins.	129
<b>Figure 5.12</b> Ratio of synonymous and nonsynonymous substitutions in <i>T. elegans</i> .	131
<b>Figure 5.13</b> Amino acid sequence logo of mature peptides in <i>T elegans</i> .	132
<b>Figure 5.14</b> Phylogeny of snakes.	136
<b>Figure 5.15</b> Blast alignment of cluster region sequences taken from between CTSB and XPO1.	137
<b>Figure 5.16</b> Conservation of synteny between snake clusters.	138
<b>Figure 5.17</b> Multiple cluster region alignment of DNA sequences from CTBS to XPO1 in snakes clusters.	139
<b>Figure 6.1</b> Genomic organisation of the <i>C. mydas</i> Beta-defensin cluster.	144
<b>Figure 6.2</b> Multiple sequence alignment of <i>C. mydas</i> beta-defensins cluster.	146
<b>Figure 6.3</b> Ratio of synonymous and nonsynonymous substitutions in <i>C. mydas</i> .	147

<b>Figure 6.4</b> Amino acid sequence Logo of second exons in <i>C. mydas</i> .	147
<b>Figure 6.5</b> Genomic organisation of the <i>G. evgoodei</i> Beta-defensin cluster.	152
<b>Figure 6.6</b> Multiple sequence alignment of <i>G. evgoodei</i> beta-defensins cluster.	153
<b>Figure 6.7</b> Ratio of synonymous and nonsynonymous substitutions in <i>G. evgoodei</i> .	155
<b>Figure 6.8</b> Amino acid sequence Logo of the second exon in <i>G. evgoodei</i> .	156
<b>Figure 6.9</b> Relationship of Genomic organisation, regions of high duplication and Phylogeny in <i>G. evgoodei</i> .	158
<b>Figure 6.1</b> Conservation of Synteny between <i>Testudine</i> Cluster regions.	159
<b>Figure 7.1</b> Agarose Gel Electrophoresis of PCR gene product and Plasmid Clone.	162
<b>Figure 7.2</b> DNA Sequences and sizing of gene insert and Plasmid MCS PCR products.	163
<b>Figure 7.3</b> Colony PCR of inserted clone into the MCS.	164
<b>Figure 7.4</b> Sequence Electrophoretogram of MSBD1 insert.	166
<b>Figure 7.5</b> Genetic translation map of sequenced region of plasmid.	167
<b>Figure 7.6</b> Gene expression and solubility testing of MSBD1.	169
<b>Figure 7.7</b> Initial antimicrobial testing of <i>E. coli</i> DH1 $\alpha$ /MBP-MSBD1 lysate.	170
<b>Figure 7.8</b> Affinity chromatography of MBP-MSBD1 UV absorbance trace.	171
<b>Figure 7.9</b> SDS PAGE gel electrophoresis of Affinity Chromatography.	172
<b>Figure 7.10</b> Antimicrobial testing of purified MBP-MSBD1.	172
<b>Figure 7.11</b> Cleavage profile of MBP-MSBD1 with TEVp.	174
<b>Figure 7.12</b> Antimicrobial testing of cleaved MSBD1 mixture.	174
<b>Figure 7.13</b> IMAC purification chromatogram.	175

<b>Figure 7.14</b> SDS-PAGE of fractions collected during IMAC chromatography.	177
<b>Figure 7.15</b> Chromatogram of Refolding Size Exclusion chromatography.	178
<b>Figure 7.16</b> 20% SDS-PAGE gel showing the fractions from the SEC chromatogram.	179
<b>Figure 7.17</b> Antimicrobial testing of SEC pooled and concentrated fractions.	180
<b>Figure 9.1</b> Schematic of the RF cloning protocol.	201
<b>Figure 9.2</b> Translation map of sequences for RF Cloning.	202

## List of Tables

<b>Table 1.1</b> Studies performed using Beta-defensin peptides showing activity levels and assays used.	25
<b>Table 2.1.</b> Positions of potential matches against tBLASTn searches using the Chicken and Green Anole Lizard.	42
<b>Table 2.2.</b> Full list of matches/predictions from BLAST searching and Gene Prediction software.	44
<b>Table 2.3</b> Comparisons of <i>C. porosus</i> sequences showing differences identified in Santana <i>et al.</i> 2021.	55
<b>Table 2.4</b> Comparisons of <i>A. mississippiensis</i> sequences showing differences identified in Santana <i>et al.</i> 2021.	56
<b>Table 3.1</b> Charge differences between longer pro-domain/mature peptides and the second exon for <i>C. porosus</i> .	58
<b>Table 3.2</b> Repeat masker summary for <i>C. porosus</i>	62
<b>Table 3.4</b> Charge differences between the longer pro-domain/mature peptides and the second exon for <i>A. mississippiensis</i> .	66
<b>Table 3.5</b> Repeat masker summary for <i>A. mississippiensis</i> .	69
<b>Table 4.1</b> Charge differences between the longer pro-domain/mature peptides and the second exon in <i>P. muralis</i> .	81
<b>Table 4.2</b> Repeat masker summary for <i>P. muralis</i> .	86
<b>Table 4.3</b> Charge differences between the longer pro-domain/mature peptides and the second exon in <i>L. agilis</i> .	90
<b>Table 4.4</b> Repeat masker summary in <i>L. agilis</i> .	95
<b>Table 4.5</b> Charge differences between the longer pro-domain/mature peptides and the second exon in <i>Z. vivipara</i> .	99
<b>Table 4.6</b> Repeat masker summary in <i>Z. vivipara</i> .	103
<b>Table 5.1</b> Repeat masker summary for <i>C. v. viridis</i> .	115
<b>Table 5.2</b> Repeat masker summary for <i>Naja naja</i> .	123
<b>Table 5.3</b> Signal peptide cleavage sites in the Beta-defensin cluster for <i>T. elegans</i> .	130

<b>Table 5.4</b> Repeat masker summary in <i>T. elegans</i> .	133
<b>Table 5.5</b> Genes common in <i>Elapidae</i> and <i>Colubridae</i> species.	135
<b>Table 6.1</b> Charge differences between the longer pro-domain/mature peptides and the second exon in <i>C. mydas</i> .	145
<b>Table 6.2</b> Repeat masker summary in <i>C. mydas</i>	149
<b>Table 6.3</b> Repeat masker summary for <i>G. evgoodei</i> .	157

## List of Abbreviations

AMP – Antimicrobial peptides  
AMR – Antibiotic Resistance  
AvBD - avian beta-defensins  
BD – Big Defensin  
BIR - baculovirus inhibitor repeat  
BLAST - Basic Local Alignment Search Tool  
BLAT - BLAST-like alignment tool  
CARD - caspase recruitment domain  
C/EBP $\alpha$  - CCAAT/enhancer-binding protein  $\alpha$   
CDH - cellobiose dehydrogenase  
CITES - Convention on International Trade in Endangered Species  
CTSB - Cathepsin B  
DAMP - damage-associated molecular patterns  
DC - dendritic cells  
*DsbC* - Disulphide Bond C  
DDSA - duplication-dependant strand annealing  
*dN* - nonsynonymous substitution  
*dS* – synonymous substitution  
ECD - extracellular domain  
*E. coli* - *Escherichia coli*  
EGFR - Epidermal growth factor receptor  
EMI - European Bioinformatics Institute  
EST – Expressed Sequence Tag  
GAL – Gallinacins  
*gor* - glutathione reductase  
HBD - Human beta-defensins  
His<sub>6</sub> - poly-histidine tag  
HMM – Hidden Markov Modelling  
IMAC - immobilised metal affinity chromatography  
LINES – long interspersed Nuclear Elements  
LLR - leucine-rich repeats  
LPG2 - lipophosphoglycan 2  
MAPK - mitogen-activated protein kinase  
MBP - Maltose binding protein  
MDA5 - melanoma differentiation-associated gene  
MSA - multiple sequences alignment  
NCBI - National Centre of Biotechnology information  
NF-IL6 - nuclear factor for interleukin 6  
NF- $\kappa$ B - Nuclear Factor Kappa B  
NOD-like - Nucleotide-binding domain leucine-rich repeat  
NR - Nitrate Reductase  
WHO – World Health Organisation  
PAMP - pathogen-associated molecular patterns  
PEG - polyethylene glycol  
PGN – Peptidoglycan  
PRR - pattern recognition receptors

PYD - pyrin domain  
RIG-like - Retinoic acid inducible gene-I  
STAT - signal transducer and activator of transcription protein  
TE – Transposable Elements  
TEVp - Tobacco Etch Virus Protease  
TLR - Toll-like receptors  
TRAM2 - Translocation associated membrane protein 2  
*trxB* - Thioredoxin reductase  
SUMO - small ubiquitin-related modifier  
UTI - urinary tract infections  
XPO1 – Exportin 1



## Chapter 1 - INTRODUCTION

### **1.1 A need for discovery.**

We live in an age where antimicrobials, in particular antibiotics, save lives. However, there is a growing problem with antibiotic resistance (AMR) worldwide. The UK government has estimated that by 2050 the global cost of AMR will be up to \$100 trillion and will account for 10 million extra deaths a year. The World Health Organisation (WHO) has declared that AMR is one of the top 10 global health threats facing humanity and the misuse and overuse of antibiotics being the main drivers in the development of drug-resistant pathogens. The WHO also suggests that the pipeline for novel antimicrobials has run dry and in 2019 WHO identified 32 antibiotics in clinical development of which only six were classed as innovative (World Health organisation 2019). Thus, there is a great need for the development and discovery of novel antimicrobials.

Antimicrobial peptides (AMP) are considered endogenous antimicrobials and can be found in many plants and animals. AMPs are considered to be part of the innate immune system (van Hoek 2014). These peptides are thought provide the 'first line' of defence against infectious microbes. In beta-defensins they are generally short cationic peptides which fold to form amphiphilic structures which increases the ability to permeate microbial cell walls (Ganz 2003). Additionally, AMPs possess a broad non-specific ability to be active towards a wide variety of organisms including both gram positive and negative bacteria, viruses, fungi, and protozoa (Zasloff 2002). Therefore, AMPs are excellent candidates for development as novel therapeutic agents to complement conventional antibiotic therapy.

#### **1.2.1 Initial discovery and characterisation**

Alexander Fleming discovered lysozyme in 1922 and marked the birth of modern innate immunity. A later example of an AMP that was isolated and characterised came from the moth *Hyalophora cecropia* in 1980 (Hultmark *et al.* 1980). This discovery helped scientists understand how insects protect themselves against microbes without an adaptive immune system. In the early 1980s, AMPs were found to exist in the mammalian leukocytes

(Patterson-Delafield *et al.* 1981) and with this AMPs were not restricted to life without an adaptive immune system.

There are thought to be three different families of AMP, based in their primary amino acid sequence along with their structure (Bals 2000):

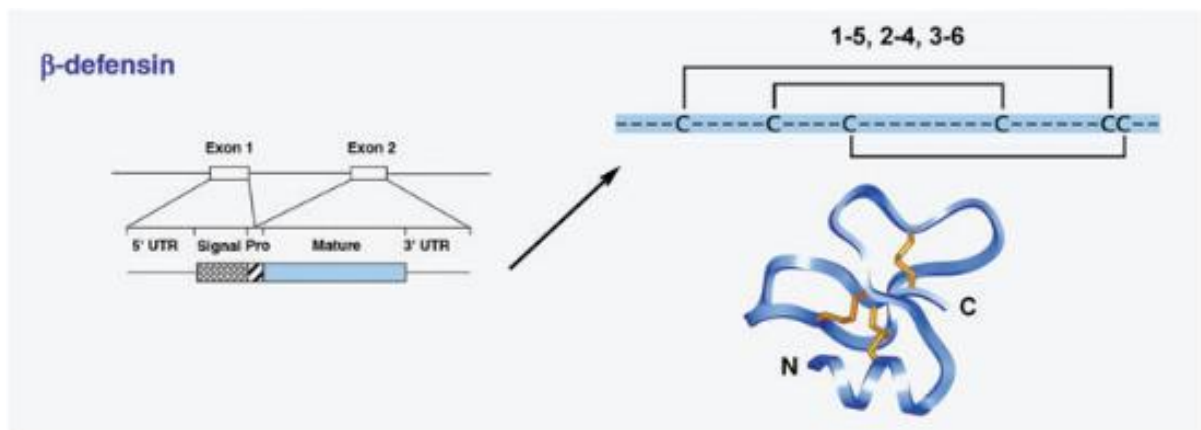
- Group 1 – Linear,  $\alpha$ - helical peptide, lacking cysteine residues for disulphide bonds (e.g., LL37 in human)
- Group 2 – Cysteine rich peptides forming disulphide bonds for stability (e.g. Defensins)
- Group 3 – Peptides containing an enrichment of one or two amino acids (e.g. histatins)

Defensins, can be put into three differing classes which are dependent on the order the cysteine bond pairings are within the mature active peptide. Defensins are one of the most common group of AMPs and there are examples of AMPs throughout the animal kingdom including arthropods, primates, birds, platypus, reptiles as well as plants, and fungi (Froy and Gurevitz 2003; Crovella *et al.* 2005; Kuo *et al.* 2015; Whittington *et al.* 2008; Stegemann *et al.* 2009; Vriens *et al.* 2014; Mygrind *et al.* 2005).

### **1.2.2 Structure and Characteristics**

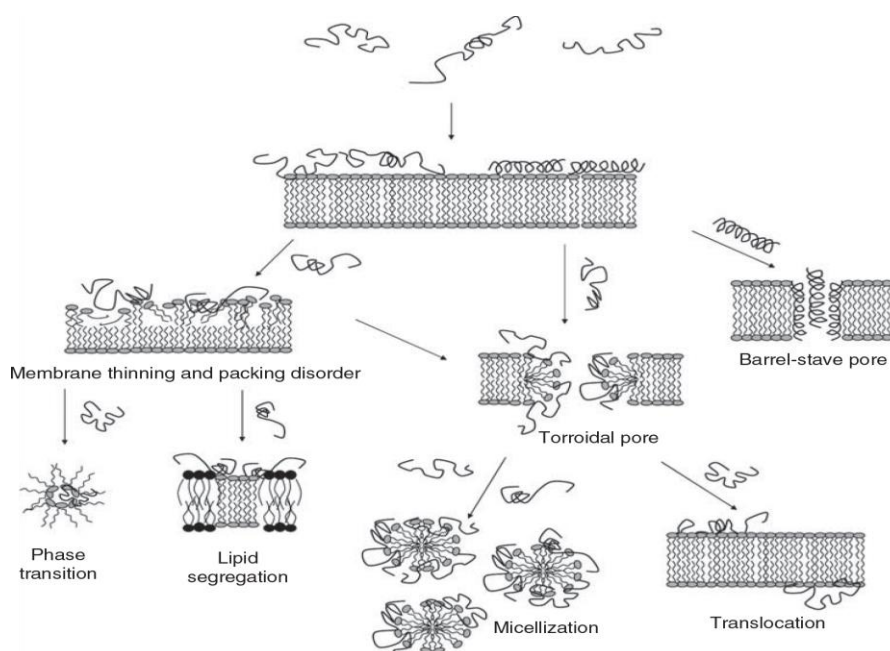
Defensins have been described as 3-4kDa cationic peptides by having 6 cysteines arranged in 3 disulphide bonds, with the characteristic pairing of the bonds highly dependent on the types of defensin (Wilson *et al.*, 2009). These peptides can also be subdivided into three group based on the order in which these bonds pair and are grouped as  $\alpha$ ,  $\beta$  and  $\theta$  Defensins, however all three classifications of defensin have only been observed in mammals (Ganz 2004). The Beta-defensin peptides are the only classification that are expressed across all the eukaryotic kingdoms (Bulet *et al.*, 2004). Defensins have mostly beta-sheet arrangements with some Alpha-helices. They also contain a large fraction of hydrophobic residues. Beta-defensins can be characterised by the conserved motif of cysteine pairing within the structure which form 3 intramolecular di-sulphide pairings between Cys1–Cys5, Cys2–Cys4, and Cys3–Cys6 as well as triple-stranded antiparallel beta-sheet configurations (Semple *et al.*, 2012).

Their exonic structure can be identified by a signal peptide followed by a pre/pro peptide finishing in a mature, active domain (Figure 1.1) (Semple *et al.* 2003; Morrison *et al.* 2003; Patil *et al.* 2004). Although AMPs induce many modes of damage to the host by hampering the cellular processes, the main action is in disruption of the cell membrane (figure2) (Malmstein, 2014).



**Figure 1.1 Beta-defensin genomic organisation.**

Signal, Pro and Mature peptide regions within exons 1 and 2 with the cysteine pairing pattern and 3D structure. (Selsted and Ouellette 2005).



**Figure 1.2. Schematic illustration of AMP interaction with lipid membranes.**

*In barrel-stave pores, peptide oligomers organise in a transmembrane structure, while toroidal pores are disorganized membrane defects caused by curvature strain. Higher peptide densities may subsequently cause complete membrane disintegration (micellisation). Furthermore, peptide binding to the polar headgroup region allows relaxation of the alkyl chains and causes membrane thinning. In addition, chemical potential gradients may result in peptide translocation across the membrane. Finally, peptide-induced lipid segregation or phase separation may contribute to AMP-induced membrane rupture (Malmsten, 2014).*

AMP's characteristics determine their ability to perform the functions demonstrated in figure 1.2. In summary these are peptide length, peptide charge, their secondary structure and hydrophobicity. There are some caveats to this. Decreased peptide length, lowers the tendency to form amphiphilic structures therefore adsorption and membrane binding decreases and therefore the efficacy of membrane lysis and antimicrobial effect (Ringstad *et al*, 2006; Deslouches *et al*, 2005). However, it has been hypothesised that if a peptide is shortened (depending on composition), this may improve performance (Sigurdadottir *et al*, 2006). In general, though, their performance lessens with decrease in peptide length. Bacterial membranes are typically anionic and are disrupted by the beta-defensin's positive charge. There is a correlation between the number of positive charges on the peptides and

the level of lysis observed, with lysis completely abolished with the removal of positive charges (Ringstad *et al*, 2007). Also, the distribution of the charges within the peptide have also been shown to play a role in the interactions of AMPs (Pasupuleti *et al*, 2012).

Formation of amphiphilic ordered structures, particularly  $\alpha$ -helices, has been found to relate to membrane disruption (Pasupuleti *et al*, 2012). An example of this is the antimicrobial protein GKE21 that has both polar/charged residues and non-polar/hydrophobic residues localised on either side of a nearly perfect  $\alpha$ -helix which results in increased membrane disruption (Pasupuleti *et al*, 2012).

In general, the more hydrophobic an AMP is the more active the peptide in membrane disruption (Aoki & Ueda 2013). However, it should be noted that increased hydrophobicity content within the amino acid sequence is strongly correlated to low selectivity and toxicity to mammalian cells (Kosikowska & Lesner 2016)

### **1.2.3 Thermal and pH stability**

Defensins and other AMPs have been shown to be remarkably stable in varying temperatures and pH ranges. A defensin-like peptide from Northwest Red beans, a commonly used cultivar of *Phaseolus vulgaris* grown in China, was shown to demonstrate antifungal activity after exposure of 100°C for 30 mins and showed resilience to pH between 0-12 (Chan & Ng 2013). Another defensin that has been investigated for its stability is a recombinant peptide from *Pseudoplectania nigrella* (Zhang *et al* 2011). In this study the defensin was shown to have antibacterial activity against *Streptococcus* and *Staphylococcus* bacteria over a pH range (2-10) and had thermostability at 100°C for 1h. An AMP isolated from *Aspergillus clavatus* ES1 has shown the promise of cysteine rich AMPs to be resistant to a range of different conditions that would normally be of detriment to peptides (Hajji *et al* 2010). The authors found that it has good activity against *Bacillus cereus* (a common bacterium involved in food borne illness) and remained stable up to 100°C and over a pH range of 3-10, however this was slightly reduced at pH 3-5 and over 10. Similar results were shown from the same peptide in another study (Skouri-Gargouri *et al* 2008). These characteristics show promise for use in a wide variety of applications in food preservation and medical or clinical settings.

### **1.3 Evolutionary origins of Beta-defensins.**

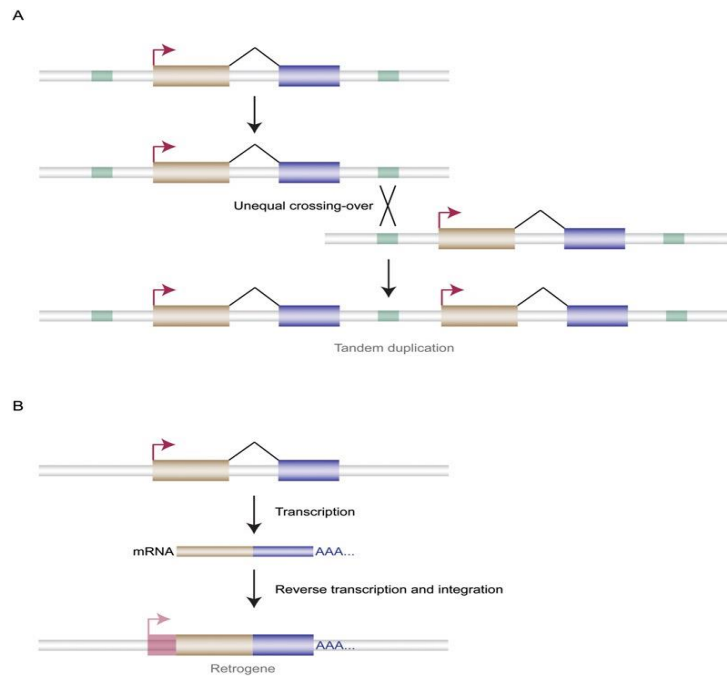
Beta-defensins as well as the other classes of defensin have been shown to have derived from a single precursor based on similarities and structures of these peptides found in other kingdoms (Semple *et al*, 2006). The identification of these defensins has led to the proposal that these peptides existed before the fungal and insect lineages of the eukaryotic domain diverged and may be at least 1 billion years old (Zhu, 2008). A conserved structural motif of two anti-parallel beta-sheets with a short turn region in between stabilised with cysteine pairings was shown across the phylogenetic spectrum (Yount *et al*, 2004) and based on its evolution it was proposed that defensins come from prokaryotic origins dating back to 2.6 billion years ago (Yount *et al*, 2006). In contrast to this, the relationship between the defensins from vertebrates and invertebrates is not so clear. There is some evidence to suggest that a group of peptides called the 'big defensins' (BDs), which have been identified in invertebrates, gave rise to the defensins found in vertebrates (Zhu *et al*, 2012). The big defensins were most recently described in amphioxus, the closest invertebrate relative to vertebrates (Teng *et al*, 2012). This enabled, with the identification in other invertebrates to trace the ancestry back to the Bilateria, indicating defensins arose around 500 million years ago (Erwin *et al*, 2002). The C-terminal region of the BDs show homology to the structure of mammalian beta-defensins (Zhu *et al*, 2012). With these identified similarities of these two groups, it suggests an evolutionary relationship between these groups of genes with the beta-defensins arising from the BDs through exon shuffling and intronisation of the sequences.

#### ***1.3.1 Gene Cluster dynamics and formation***

Beta-defensins are usually found in clusters (Patil *et al*. 2005) along a relatively small region of the genome. Clusters are usually formed from one or more parental genes duplicating to give rise to daughter genes with this process repeating many times to give rise to several genes within the cluster (Xiao *et al*. 2004). Since 1970, a book written by Susumu Ohno, *Evolution by Gene Duplication*, has suggested that gene duplications are a driver for many novel genes and therefore an important factor in driving evolution. He emphasised that a duplicate of a gene would offer new opportunities allowing the one of the gene copies to evolve possible new functions with the other copy preserving the original function. This idea

is described as neofunctionalisation. However, Ohno (1972) goes onto say that the outcome of most duplicate genes is that of pseudogenisation whereby the gene, over time, becomes redundant. Later, it was proposed that potentially multiple functions may arise from a duplication with this process being named 'subfunctionalisation' (Conant and Wolfe 2008). Rather than one single gene copy degrading or evolving a new function the duplicates perform subtly different functions and selective pressures may result in a compromise between optimal sequences for each role (Hurles, 2004). This could have a beneficial effect by removing a conflict between functions and with beta-defensins in mind, subfunctionalisation could be a valid outcome for a duplication of genes within the cluster. However, neofunctionalisation occurs where genes are rapidly evolving such as the host's defence and immunity (Emes et al. 2003).

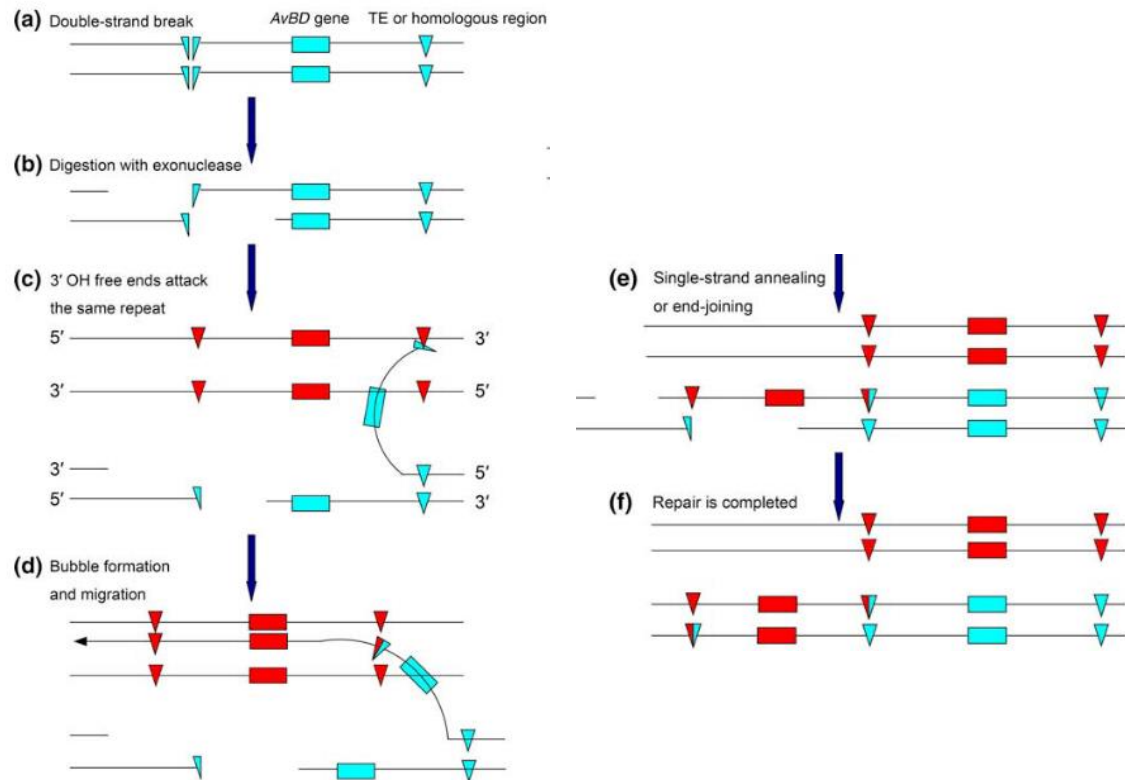
DNA duplication events occur by varying mechanisms and are classified on the size of the duplication and whether an RNA intermediate is involved. Small scale duplications involve the duplication of a single gene or a segment of chromosome (termed segmental duplications), through whole genome duplications resulting in polyploidism. RNA-based duplications that have occurred through retro-transcription through transposable elements (TEs) result in a duplication that is randomly inserted into the genome. Given that whole-genome duplication is a rare event in vertebrates (Mable 2004) and reptiles, the most likely mode of duplication in reptiles is segmental, where the highly repetitive nature of reptile genomes provides sequences of pseudohomology facilitating unequal crossover during homologous recombination in meiosis, which in turn will produce tandem duplications ultimately forming a gene cluster (figure 1.3) (Shedlock *et al* 2007). If homologous recombination were to be repeated a few times, it would ultimately give rise to many duplicate genes (Hargreaves *et al* 2018). In contrast to this, TEs have been shown to reside in the defensin region of Galliformes and Passeriformes (Chen *et al.* 2015). It was proposed that the presence of TEs within this region provide a basis for duplication-dependant strand annealing (DDSA) first proposed by Fiston-Lavier (2007). The model describes that after a double strand break the end of the sequence automatically searches for a corresponding homologous region, likely to be a close by TE with the resultant repair creating TE hybrid copies and tandem gene copies (figure 1.4). Additionally, it was found that an exceptionally high content of TEs are distributed in highly duplicated regions of the defensin gene clusters in the golden pheasant and hwamei (Chen *et al.* 2015).



**Figure 1.3. Origin of new gene copies through gene duplication.**

(A) DNA-based duplication. A common type of segmental duplication—tandem duplication—is shown. It may occur via unequal crossing-over that is mediated by transposable elements (light green). There are different fates of the resulting duplicate genes. For example, one of the duplicates may acquire new functions by evolving new expression patterns and/or novel biochemical protein or RNA functions. (Gold and blue boxes) Exons, (black connecting lines) exon splicing, (red right-angled arrows) transcriptional start sites (TSSs), (grey tubes) nonexonic chromatin. (B) RNA-based duplication (termed retroposition or retroduplication). New retroposed gene copies may arise through the reverse transcription of messenger RNAs (mRNAs) from parental source genes. Functional retrogenes with new functional properties may evolve from these copies after acquisition or evolution of promoters in their 5' flanking regions that may drive their transcription. (Pink right-angled arrow) TSS, (transparent pink box) additionally transcribed flanking sequence at the insertion site. (H. Kaessmann 2010)





**Figure 1.4. Putative mechanism for defensin gene duplication.**

Duplication occurs by nonallelic homologous recombination between transposable elements (golden pheasant) or homologous regions (hwamei). This model is modified from Fiston-Lavier et al. A) Double-strand break within a transposable element or homologous region. B) Exonuclease digests the 5' extremity and exposes 3'OH free ends. C) 3'OH free ends attack the same repeat region. D) Bubble formation and migration. E) Single strand annealing or end joining to repair the second strand. F) Finishing the repair process (Chen et al. 2015).

### 1.3.2 Selection pressures

As outlined above gene duplication is a fundamental process by which novel proteins with novel functions evolve but the mechanism for the new functions remains uncertain. There are, however, problems with this theory (Hughes 1994). There is now evidence that after gene duplication, positive Darwinian selection occurs and not the random accumulation of mutations to give rise to the daughter genes within a cluster. The analysis of synonymous (*dS*) and nonsynonymous (*dN*) nucleotide substitutions gives an insight into the evolutionary divergence of the mutations within these duplications (Hughes 1999). Synonymous or silent mutations are usually invisible to natural selection as these are mutations that do not alter

the amino acid in the protein sequence as the codons are shared (Akashi 1995), whereas, nonsynonymous mutations change the amino acid codon, which may be under greater selection pressure. Therefore, comparing these types of mutations may provide mechanisms for the evolution of duplicated genes. For example, models of variable synonymous/nonsynonymous rate ratios ( $dS/dN$ ) among these codon sites might provide an insight into the limitations among amino acids in the protein and provide which of these sites might be under positive selection within a gene (Yang and Nielson 1998). The most used approximations of these rates were devised by Nei and Gojobori (1986).

In relation to the evolution of beta-defensins a few notable papers have been published. The evolution of avian beta-defensin genes were investigated (Chen *et al.* 2015) and the findings suggested that there were relatively low  $dS/dN$  ratios within the defensin sequence that is indicative of purifying selection due to the possible constraints of the tertiary structure to provide a defence against microbes. However, when looking at the difference in these ratios there was a greater  $dS/dN$  ratio in the mature active peptide compared to the signal peptide, suggesting that the 'arms race' between pathogen and mature peptide is providing an intense selection pressure in this part of the peptide. This was also conferred by analysis of the similar alpha-defensin gene clusters of rodents and primates (Patil *et al.* 2004). Sites within the defensin gene show that within the characteristic defensin beta-sheet these are largely unaffected by positive selection as these form the basis of oligomerisation of beta-defensins, however, there was a greater degree of positive selection within the alpha-helices of the N-terminal portion of the peptide. Since alpha-helices are often associated within membranes it is speculated that these are involved with the anchoring to the pathogen, therefore playing a large role in their AMP properties (Semple *et al.* 2006).

#### **1.4 Role of Beta-defensins in innate immunity.**

Initiation of innate immunity involves an important cooperation between three different sensing receptors – Toll-like receptors (TLRs), nucleotide-binding domain leucine-rich repeat (NOD-like) receptors (NLRs) and retinoic acid inducible gene-I (RIG-like) receptors (RLRs) (Creagh and O'Neill 2006). These receptors are collectively known as pattern recognition receptors (PRRs), and they recognise specific molecular signatures known as damage-associated molecular patterns (DAMPs) (Creagh and O'Neill 2006) and pathogen-associated

molecular patterns (PAMPs) (Kawasaki and Kawai 2014). PAMPs are characteristically conserved molecules within a class of microorganisms, such as lipopolysaccharide (LPS) bacterial DNA (unmethylated CpG DNA),  $\beta$ -glucans, flagellins and peptidoglycans (Contreras, G. *et al.* 2020).

#### *1.4.1 Toll-like Receptors – Membrane Bound Sensors*

Toll-Like receptors are transmembrane glycoproteins and highly conserved between vertebrates (Leulier and Lemaitre 2008). Each TLR gene consists of an intracellular (cytoplasmic) TIR domain, responsible for cellular signalling (Medzhitov *et al.* 1997), a conserved transmembrane region and a variable extracellular domain (ECD) involved in the ligand recognition of PAMPs. Each ECD consists of varying numbers of leucine-rich repeats (LLRs) motifs (Matsushima *et al.* 2007). These are located on the cell surface or on endosomes.

TLRs are expressed in innate immune cells such as dendritic cells (DCs) and macrophages as well as non-immune cells such as fibroblasts and epithelial cells (Kawasaki and Kawai 2014). TLRs are thought to be divided into two distinct subclasses according to the ligands that they bind (Wlasiuk and Nachman 2010). TLRs 1,2,4,5,6 and 11 are expressed on the cell membrane and detect PAMPs associated with bacterial components and TLRs 3,7,8 and 9 are intracellular and detect ssRNA and dsRNA that are associated with viral infection (Akira *et al.* 2006).

#### *1.4.2 NOD-like receptors – Intracellular Cytoplasmic sensors*

NOD-like receptors (NLRs) are a family of cytoplasmic pathogenic sensing proteins which are conserved among members of the plant and animal kingdom responsible for playing varying roles in inflammation, apoptosis, and host defence mechanisms (Ausubel 2005). NLRs are generally expressed in immune cells such as lymphocytes and antigen-presenting cells (i.e. macrophages and dendritic cells). They are also expressed in non-immune cells including epithelial cells and primarily recognise bacterial PAMPs and endogenous danger signals (Lee and Kim 2007). NLRs structure consists of three differing domains: 1) a variable N-terminal protein-protein interaction domain, being either, a caspase recruitment domain (CARD), a pyrin domain (PYD), acidic transactivating domain or baculovirus inhibitor repeat (BIR); 2) a

central oligomerisation (NOD) domain which is responsible for self-oligomerisation when activated (Inohara, N. *et al.* (2000); and 3) a C-terminal LLR which detects PAMPs. The N-terminal domain is responsible for the downstream signalling cascades. CARD domains were first thought to be associated with apoptosis and inflammation through caspases, but CARDS have also been shown to mediate caspase-independent interactions (Chen *et al.* 2009). The structure of PYD is homologous to CARD and are both members of the death-domain superfamily involved in apoptosis and inflammation. Lastly, the BIR-containing proteins are classed into two groups, the neuronal apoptosis inhibitor proteins (NAIPs) and the inhibitor of apoptosis proteins (IAPs). The NLRs that have been identified in reptiles are NOD1, NLRC3 and NLRX1 and interestingly NOD2 was not found present in the genomes in this study (Chen, J. *et al.* 2019) and therefore will not be in the scope of this review. The agonists for NOD1 have been identified as specific amino acids of Peptidoglycan (PGN) (Uehara *et al.* 2006), NLRC3 has been shown to regulate the inflammatory response (Schneider *et al.* 2013) and NLRX1 also acts as a checkpoint for overactive inflammation (Allen *et al.* 2012).

Signalling in the NLRs is very similar to that of TLRs which share downstream targets. NLRs activate signalling pathways for pro-inflammatory mediators to defend against infection. Along with TLRs, NLRs activate Nuclear Factor Kappa B (NF- $\kappa$ B) and mitogen-activated protein kinase (MAPKs) (Girardin *et al.* 2001). Upon NOD stimulation it self-oligomerises to recruit RICK (Also RIP2) which is essential for the activation of both NF- $\kappa$ B and MAPKs (Inohara *et al.* 2000).

NOD1 is expressed in epithelial cells and activates chemokines for the recruitment of immune cells and importantly the production of antimicrobial peptides, such as, beta-defensins showing a specific role in epithelial innate immune protection. It has also been shown that NLRs interact synergistically with TLRs in response to PGN (Uehara *et al.* 2005) however, NLRs may also provide a backup defence mechanism when TLR signalling has become tolerant to certain stimuli.

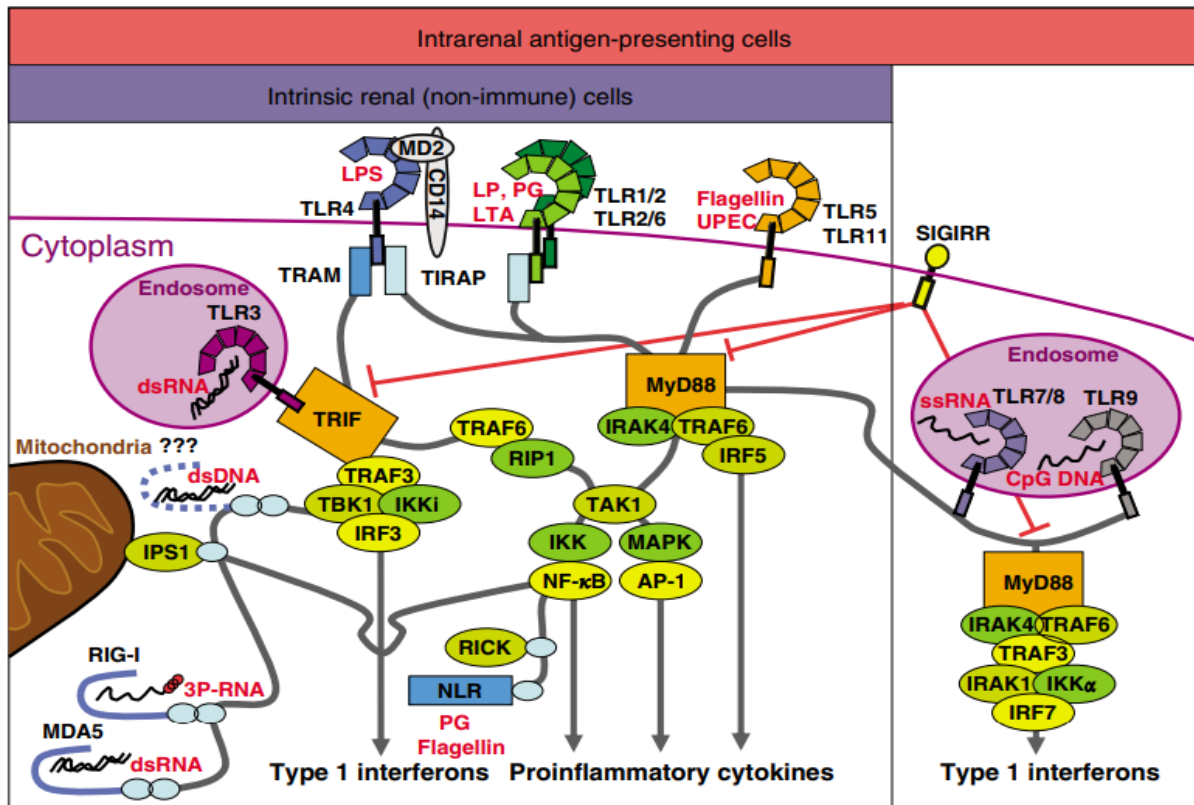
#### 1.4.3 RIG-like receptors – intracellular viral sensors

The TLRs responsible for viral protection are localised in endosomes and help with downstream signalling once the virus has been internalised and lysed to release the nucleic acids, however, once a virus has directly entered the cytoplasm and released viral-dsRNA,

TLRs can no longer recognise this. This is where cytoplasmic sensors in the host cell come into play to detect actively replicating viruses, namely the RIG-like receptors (RLRs). The RLRs belong to a family of DExD/H box RNA helicases and comprise of 3 members – retinoic acid - inducible gene (RIG-1), melanoma differentiation-associated gene (MDA5) and lipophosphoglycan 2 (LPG2) (Yoneyama and Fujita 2009). These three genes have been identified in reptile genomes (Chen *et al.* 2019).

Human beta-defensins (HBDs) are the most studied of the mammalian defensins and are primarily expressed in various epithelial tissues, along with some immune cells, such as monocytes and macrophages (Phoenix *et al.*, 2013). The most well studied and known roles of the vertebrate beta-defensins are that they are part of the innate immune system with an effective action against a whole range of different pathogens. In addition to this they play a role in the protective response to infection by having a critical role in regulating the inflammatory response (Semple *et al.*, 2012) and have also been shown to function as a chemoattractant (Yang *et al.*, 1999).

Figure 1.6 summarises the cellular pathways in which these receptors play in signalling and activation of transcription factors involved in gene expression associated with the innate immune system in renal cells.

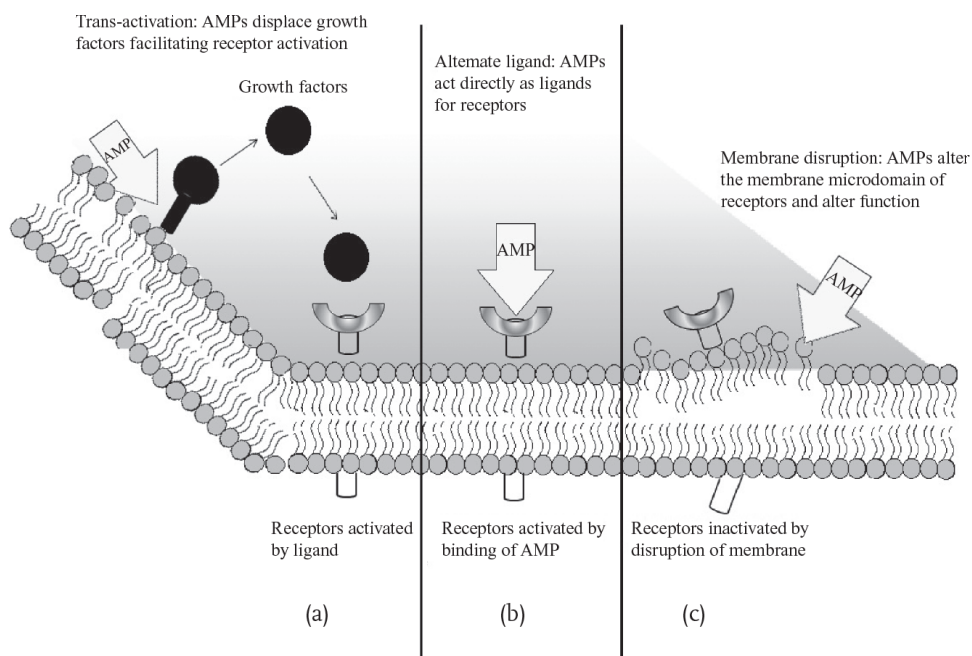


**Figure 1.6. TLRs, RLHs, and NLRs in renal cells.**

*Intrinsic renal non-immune cells express a restricted TLR pattern as compared with intrarenal myeloid dendritic cells (and macrophages). Whether RLHs and NLRs are expressed by both types of renal cells has not yet been characterized in detail. TLRs signal through the MyD88 and/or TIR domain-containing adaptor protein-inducing-interferon- $\beta$ -signalling pathways for the induction of proinflammatory cytokines and type I IFNs. The RLHs signal through mitochondrial interferon- $\beta$  promoter stimulator-1 and NLR use receptor-interacting protein-like interacting CLARP kinase. The respective ligands to these receptors are indicated in red (LPS, lipopolysaccharide; PG, peptidoglycan; LTA, lipoteichoic acid; LP, lipoprotein; 3P-RNA, 50-triphosphate RNA) (Anders 2007)*

It was shown that recruitment of immature dendritic cells and CD4<sup>+</sup> memory T cells occurs with a concentration gradient of HBDs 1 and 2. HBD3 also possesses chemoattractant properties towards immature murine dendritic cells showing a positive relationship between innate and adaptive responses in mammals, however, the flip side of this is that the mechanisms by which beta-defensins mediate chemotaxis are poorly understood (Hazlett *et al*, 2011). However, three modes have been proposed (Figure 1.7) (Lai *et al*, 2009). The

‘transactivation model’ suggests that Beta-defensin stimulates the release of a membrane bound growth factor which in turn binds to the receptor, the ‘alternate ligand model’ shows that the beta-defensin binds directly to the receptor initiating signalling and the ‘membrane disruption model’ suggests that this causes a signal to initiate or to become unresponsive. It has been suggested that Beta-defensins may also play a role in suppression of a proinflammatory response (Semple *et al*, 2012) however these remain unclear. The evidence outlined indicates that beta-defensins may combine pro/anti-inflammatory responses by balancing these effects through the expression levels of the peptides.



**Figure 1.7. Models for the interaction of beta-defensins with cell surface receptors:**

*Putative models for the interaction of defensins with chemokine receptors. According to the “trans-activation model”. A) AMPs stimulate the release of a membrane-bound growth factor, which then binds to its high-affinity receptor with activation resulting. In the “alternate ligand model”. B) AMPs bind directly to the receptor, which results in the initiation of signalling. The “membrane disruption model”. C) proposes that AMPs modify the membrane microdomain associated with the receptor, which indirectly leads to a change in receptor function. This functional change allows the receptor to either signal without a ligand or become insensitive to binding by its specific ligand (Phoenix *et al*, 2013)*

#### 1.4.4 Beta-defensins and Pattern Recognition Receptors (PRR) recognition and activation.

Epithelial cells present the first cell surface to encounter potential pathogens. Various studies have, with the knowledge set out above, explored the inducibility and regulation of beta-defensins. Whilst some beta-defensins may be constitutively expressed in some tissues, their expression can be upregulated in other tissues in response to microbial infection (van Dijk 2008). Human beta-defensin 2 (hBD2) has been shown to be stimulated in response to LPS (Tsutsumi-ishi and Nagaoka 2001). In this study they explored the roles of NF- $\kappa$ B, signal transducer and activator of transcription protein (STAT) and a nuclear factor for interleukin 6 (NF-IL6) in transcriptional regulation of the release of hBD2 using a luciferase reporter assay. It was found that 2 NF- $\kappa$ B sites were crucial for basal transcriptional activity and neither NF- $\kappa$ B, Interleukin 6 (IL6) or STAT was required for the induction of hBD2. This suggests that PRRs could be responsible for the release of inducible beta-defensins.

Transcriptional regulation of beta-defensins has also been shown to be induced in a similar manner in tracheal epithelial cells indicating a link between PRRs and the importance of this in epithelial surfaces (Diamond *et al.* 2000). The authors also suggest that antimicrobial peptide based innate immunity is conserved among evolutionarily diverse organisms. Oesophageal cells, when challenged with *Candida albicans*, have been shown to activate NF- $\kappa$ B and AP-1 and inhibition of these pathways revealed that hBD2 is synergistically regulated by these factors (Steubesand *et al.* 2009). This study also gave insight into hBD3 being independently regulated, not by NF- $\kappa$ B, but solely on Epidermal growth factor receptor (EGFR)/MAPK/AP-1- dependent pathway. Linking the potential of PRRs being involved in the regulation of beta-defensins and their importance in being a critical part of host defence at mucosal surfaces *In vitro* studies have demonstrated that NOD1 is stimulated by *Bacillus* and *Shigella* by release of PGN fragments in a NF- $\kappa$ B dependant manner (Hasegawa *et al.* 2006, Nigro *et al.* 2008).

There is mounting evidence that there is a close relationship between TLRs and beta-defensins, especially extracellular TLRs. Studies have shown that murine BD2 (mBD2) and LPS share signalling pathways through the TLR4 receptor (da Silva Correia *et al.* 2001) and mBD2 can activate NF- $\kappa$ B in human embryonic kidney cells (HEK) 293 cells transfected with TLR4 and mBD2, more specifically mBD2 acted as an endogenous stimulator for TLR4 inducing maturation of DCs (Biragyn *et al.* 2002). It has been reported that TLR2 and TLR4 in human



cholangiocytes can differentially regulate the production of hBDs promoting host epithelial resistance to *Cryptosporidium parvum* (Chen *et al* 2005). Also, it has been observed that activation of professional APCs by hBD3 is mediated by interaction with TLR1 and 2 through the Myeloid differentiation primary response 88 (myD88)/ Interleukin-1 receptor-associated kinase (IRAK)/NF- $\kappa$ B pathway (Funderberg *et al* 2007).

Beta-defensin promoter regions have been described by Yang and Oppenheim (2003). They contain CCAAT/enhancer-binding protein  $\alpha$  (C/EBP $\alpha$ ), NF-IL6, activation protein-1 (AP-1) which could play a part in maintaining constitutive expression and NF- $\kappa$ B, and interferon- $\gamma$  (IFN- $\gamma$ )-activated site (GAS) which could be inducible in response to stimuli. Similar findings were found in chicken beta-defensin 9 promoter (Van Dijk *et al.* 2007).

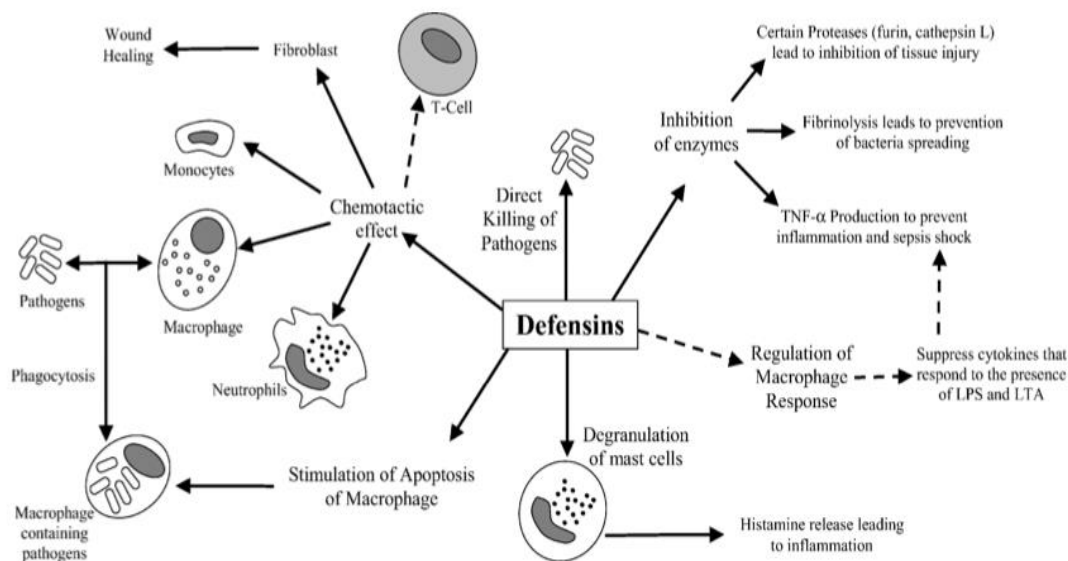
These investigations, addressing PRR recognition and signal transduction can detail the earliest immunological events following infection (Zimmerman 2020), for example, defensins may be expressed at high levels at the site of pathogen entry resulting in an inflammatory response which may recruit other cells of the immune system (Semple and Dorin 2012). Oppenheim and Yang (2005) demonstrated the chemoattraction of immature DCs and McDermott (2004) described HBD-1 are chemotactic for monocytes and HBD2 and HBD3-4 recruit mast cells and macrophages. As these pathways are highly conserved among species this could help in disseminating the gaps within reptilian lineages. For example, NF- $\kappa$ B translocates to the nucleus following infection in crocodiles that is comparable to human mechanisms (Merchant *et al.* (2017). In turtles signalling through TLR4, stimulated by LPS, upregulates IL-1 $\beta$  which involves the NF- $\kappa$ B pathway (Zhou *et al.* 2016).

#### 1.4.5 Wound healing

Turning to examine wound healing, when the natural barrier of the skin is damaged and is open to infection, DAMPS and PAMPS play a role in how the body responds to potential pathogens. With the downstream signalling of the innate immune response known this can then be targeted with potential therapeutics such as AMPs in particular beta-defensins. One example of where these potential novel therapeutics is with diabetic patients where it is known they have a particularly hard time in healing infected wounds. Research suggests that beta-defensins may be used directly against bacterial and viral infections to promote wound

healing and it has been shown that Human beta-defensin (hBD)-3 expression significantly promotes wound closure in diabetic pigs (Hirsch *et al*, 2009). These findings reveal that hBD-3 may play a major role in diabetic wound healing. From this it can be deduced that beta-defensins have a large part to play in the resistance to infection in wound healing and further studies should be done in this area to understand the mode of action of these peptides. A study has also shown that in a high glucose environment the expression of HBD2 is downregulated which in turn impairs keratinocyte migration and new capillary formation at the site of the diabetic wound (Baroni *et al*, 2009). Growth factors such as EGF and Insulin growth factor-1 have been shown to induce expression of HBD3 in keratinocytes (Sorenson *et al*, 2003) (Lan *et al*, 2011). However, this study also describes that in a diabetic rat model, rat beta-defensin 3 (HBD3) expression was negligible under hyperglycaemic conditions. This was consistent with HBD-3 expression in human studies.

With respect to wound healing in reptiles, one such example has been studied in the Anole Lizard (*Anolis carolinensis*) (Alibardi *et al*, 2012). This study looked at where previously identified beta-defensins originated (Dalla Valle *et al*, 2011). The lizards resist infection after tail loss and the team investigated the injured tissues of the lizards after tail loss. They found that within the first week post injury, granulocytes and keratinocytes were present and through further investigation it was found that the granulocytes present in the wound contained dense non-specific (azurophil) granules present in mammalian granulocytes. It was hypothesised that the resistance was due the high numbers of granulocytes at the site of the wound, and it showed that a beta-defensin was present whereas it was not in healthy tissue. This once again reinforces that beta-defensins play a major role in the effectiveness of wound healing. Figure 1.8 summarises how Defensins are associated with inflammatory and defensive responses.



**Figure 1.8. Principal functions of defensins in inflammatory and defensive reactions against pathogens** (Sugalaro et al. 2004).

### **1.5 Beta-defensin Discovery and experimental research methods**

With extensive review of the current state of research into beta-defensins and given that antibiotic resistance has become an increasing problem in recent years, partially because antibiotic use is rising but also because the pace at which we are discovering novel antibiotics has slowed drastically, antimicrobial peptides could offer a new and novel therapeutic weapon in the fight against infection. A summary of some of the current experimental research methodology in Beta-defensin research will be outlined below:

#### **1.5.1 Data mining approach**

Given the highly conserved nature of beta-defensins, it is possible to use a data mining approach to the discovery of novel beta-defensins in other species. With the advent of faster sequencing there are many sequences yet to be investigated. The main source for these sequences is web-based databases such as the different libraries in the National Centre of Biotechnology information (NCBI) (<http://www.ncbi.nlm.nih.gov>). This was shown in the discovery of beta-defensins like homologues in *Pelodiscus sinensis* (Yu et al. 2016). In their

study, they discovered several beta-defensins by using a program (<http://www.ncbi.nlm.nih.gov/blast>) which looks for similarities between known and unknown sequences. With *Pelodiscus sinensis*' evolutionary relationship with the green lizard and the chicken, Yu *et al.* could identify potential beta-defensins in *Pelodiscus sinensis*. This tool has been used in many other gene discovery studies and therefore will play a pivotal role in this research (Li *et al.* 2003, Schutte *et al.* 2002, Zou *et al.* 2007, Lynn *et al.* 2004, Jalkanen *et al.* 2005).

Over the years researchers have employed several different methods to identify and characterise beta-defensins. The chicken beta-defensins have been the most characterised out of the avian beta-defensins. The first two reports of avian beta-defensins, which were named Gallinacins, were isolated from chicken leukocytes, and named GAL-1, 1 $\alpha$  and 2 (Harwig *et al.* 1994; Evans *et al.* 1995). In these studies, the peptides were extracted and purified from the separated granules of the leukocyte. These peptides were sequenced using LC-electrospray mass spectrometry. Later, in 2004, a fourth Gallinacin, named GAL-3 was characterised (Zhao *et al.* 2001) through sequencing of cDNA clones extracted from epithelial cells. These techniques of gene discovery are labour intensive and do not provide information on their genomic sequence or location.

From the absence of the ability to sequence the cDNA obtained from tissues, through to the advent of sequencing data, such as Expressed Sequence Tags, and subsequently sequencing of genomes, *in silico* methods of identification have been explored.

Since 2004, nine more novel beta-defensins have been identified in chicken using bioinformatic techniques, notably using EST libraries that were publicly available (Lynn *et al.* 2004). An Expressed Sequence Tag, a short sub-sequence of a cDNA sequence is used to identify gene transcripts, however with different *in-silico* technique, Hidden Markov Modelling (HMM) other beta-defensins have been identified (Santana *et al.* 2021). This was because the beta-defensins share a conserved motif making it possible to produce an HMM profile. An HMM profile is a probabilistic model, which uses position specific scores to indicate the likelihood of each amino acid occurring in each position in an alignment (Eddy 1998). Two more Chicken beta-defensins were identified in 2005 (Higgs *et al.* 2005) and more

recently several crocodilian beta-defensin clusters were identified using this method (Santana *et al.* 2021).

In March 2004 the chicken genome was sequenced by the International Chicken Genome Sequencing Consortium, which then allowed a Genome wide screen to identify a single Beta-defensin cluster (Xiao *et al.* 2004). In the study, all known defensin-like peptide sequences were individually queried against the chicken nonredundant (NR), high throughput genomes sequences (HTGS) and whole-genome shotgun sequences (WGS) databases in the GenBank by using the Basic Local Alignment Search Tool (tBLASTn) program (Altschul *et al.* 1990). Once all the putative defensin genes were retrieved the defensin cluster was generated into a longer contig. Later this was confirmed using BLAST-like alignment tool (BLAT) (Kent 2002) with the release of the chicken genome.

The beta-defensins discovered in these studies were compiled into standard nomenclature (Lynn *et al.* 2007). These were designated the term “avian beta-defensins” (abbreviated to AvBD) to describe this group of molecules.

In 2014, in a paper about antimicrobial peptides in reptiles, Van Hoek noted which beta-defensins had previously been described. The two species that had several genes identified were the Red-eared Slider Turtle (Kaplinsky *et al.* 2013) in which the transcriptome had been sequenced and from this several genes identified, and the Green Anole Lizard (Dalla Valle *et al.* 2012) whereby the genes identified were deduced from BLAST searching against EST libraries. In both cases chicken beta-defensins were used for the prediction and comparison of the beta-defensins found.

In 2017, the first crocodilian cluster was determined (Tang *et al.* 2018). In this study they retrieved all known reptile and avian beta-defensins from the NCBI databases (<http://www.ncbi.nlm.nih.gov/>) and queried these individually against the Chinese Alligator genome (Wan *et al.* 2013).

### 1.5.2 Sequence alignment and Phylogeny Analysis.

Sequence alignment is a way of arranging the sequences of either DNA, RNA, or protein to identify regions of similarity and consequently may show regions that are functionally, structurally, or evolutionarily similar (Mount, 2004). The aligned sequences are usually represented in rows with gaps being inserted between identical residues so that identical or

similar sequences can be put into columns. There have been many studies that have used this method to show relationships between similar sequences, however, these sequences are usually done using multiple sequences alignment (MSA) whereby three or more sequences are aligned using software packages that utilise algorithms. Most notable is CLUSTAL and its series of later iterations and is now the standard tool for this. Investigations into novel and discovered reptile and fish beta-defensins have used this tool to show similarities within species as well as between species and to show evolutionary relationships (Zou *et al.*, 2007; Soman, Arathy and Sreekumar, 2009; Correa and Oguiura, 2013).

### 1.5.3 Protein expression and purification.

Recombinant DNA technology offers more sustainable, scalable, and cost-effective means to produce AMPs as genetically modifying micro-organisms have been used flexibly for many years and does not cause the issues that producing larger peptides can have when using chemical synthesis methods for protein synthesis (Müller 2015). Once the AMP gene coding regions are identified in genomic studies these regions can then be amplified by the polymerase chain reaction (PCR) in order to be cloned into a suitable vector, such as a plasmid, and expressed in a suitable expression host to then be processed further downstream, purified, and studied.

As beta-defensins derive from eukaryotes their genes have multiple coding and non-coding regions and because of their relatively simple cellular machinery, bacteria lack the ability to accomplish proper post-translational modifications and molecular folding needed to express a fully functional protein. Eukaryotic systems have been employed to overcome this; however, many studies have used bacterial cell lines for cloning and expression. *Pichia pastoris* has been used in several studies to express AMPs (Yu *et al.*, 2016; Zhang *et al.*, 2011). *P. pastoris* is a methylotrophic yeast which has a high growth rate and can be grown inexpensively in a simple medium. There are many different expression vectors available with one example study by Yu *et al.* (2011) used a recombinant pPIC9K vector, pPICPs-BD2. This contained the full *Pelodiscus sinensis* BD2 (Ps-BD2) open reading frame (ORF). The vector was constructed by fusing the ORF directly with an EcoRI site and his-tag upstream along with a NotI site downstream. A Tobacco Etch Virus Protease (TEVp) site was incorporated between the ORF and poly his-tag. HEK293T cells have also been used a vector for expressing beta-

defensins (Soman *et al*, 2009). Another expression vector that has been shown to produce defensins with some success has been using a Nitrate Reductase (NR)- deficient *Chlorella ellipsoidea* mutant nrm-4 (Bai *et al* 2013). In this study a plant expression plasmid vector was created to contain the defensin NP-1 gene from rabbits. It was transformed into the *C. ellipsoidea* by electroporation. Once a transgenic line had been formed, they found that it had produced defensins at high levels at approximately 11.42mg/l. As the *C. ellipsoidea* was NR-deficient, isolation of the transgenic strain was cultured under selective medium to allow for extraction and purification.

*Escherichia coli* (*E. coli*) is by far the most used microbial system to produce recombinant AMPs (Wibowo *et al*. 2019). Its genetics are well documented, and its ease of manipulation owes itself to a good candidate when it comes to recombinant expression. The strain BL21(DE3) is commonly used as it lacks ompT and Lon proteases which may inhibit the production of AMPs to sufficient levels (Sørensen and Mortensen 2005). One drawback of using BL21(DE3) is that it fails to express fusion AMPs which contain disulphide bonds such as Defensins and instead formed protein aggregates, which are undesirable for downstream processing (Panteleev and Ovchinnikova 2017). One such way around this is to co-express the signal sequence from *MalE* which directs the fusion protein to the periplasm where the cellular machinery can provide the oxidation environment for the cysteine bonds (Klint *et al*. 2013). Some strains of *E. coli*, Origami and Rosetta-gami, have been specifically designed and developed to provide an ability to form disulphide bridges in the cytoplasm by disrupting both the Thioredoxin reductase (*trxB*) and glutathione reductase (*gor*) pathways. The *trxB* and *gor* genes provide a reducing environment in the cytoplasm which inhibits the formation of disulphide bridges. The downside of these strains is that the production of multiple disulphide-bonded proteins can be misoxidised and stay this way due to the lack of isomerisation. This is due to the lack of the periplasmic protein Disulphide Bond C (*DsbC*) which provide the correct folding and disulphide bond formation. SHuffle® (New England Biolabs) is a strain which expresses cytoplasmic DsbC which provides isomerase activity for the correct formation and bonding within the protein. This was successfully demonstrated in the expression of membrane protein U24 from human herpes virus (Tait and Straus 2011). By their very nature AMPs have toxicity to expression hosts, which in turn, make them difficult to express on their own. Fusion tags have become a further technique employed to overcome these challenges. In addition to providing a means of reducing the overall charge and

increasing their molecular weight, fusion tags can provide an effective purification strategy. Cleavage sites are also incorporated into the constructs which allow the liberation of the AMP after purification. The fusion tag small ubiquitin-related modifier (SUMO) is a 11.60kDa, 101 amino acid protein, with a poly-histidine tag (His<sub>6</sub>) to facilitate purification with immobilised metal affinity chromatography (IMAC) has been widely used (Lin *et al.* 2017; Luan *et al.* 2014; Wei *et al.* 2018). Using SUMO also has no need to provide a cleavage site as this is recognised by SUMO protease and the solubility of the fusion AMP is also enhanced through the hydrophilic outer and hydrophobic core structure (Yadav *et al.* 2016). Like SUMO thioredoxin A (Trx), a 11.68kDa, 109 amino acid protein tag is also used in combination with His<sub>6</sub> at the C-terminus for IMAC purification as well as a specific sequence at the N-terminus for enzymatic cleavage, such as tobacco etch virus protease (TEVp) (Herbal *et al.* 2015; Li 2013). Maltose binding protein (MBP) is a fusion tag which has a dual function as an enhancer for solubility and is an affinity tag for purification by chromatography using an amylose resin (Li *et al.* 2014; Vu *et al.* 2014). A His<sub>6</sub> tag is also added to the N-terminus to provide another layer of purification method by IMAC. In conjunction with a TEVp site is placed between the AMP and MBP to allow the cleavage of the AMP. New England Biolabs also produce a TEVp with His<sub>6</sub> tag so after purification of the fusion protein has been performed and TEVp used to cleave, IMAC can be used to capture the cleaved fusion and TEVp allowing the AMP to be captured in the flowthrough. In addition to this, Yu *et al.* (2016) used Ni-NTA affinity chromatography as a first step of purification and gel permeation or size exclusion chromatography has also been used for purification (Mygind *et al.*, 2005).

#### *1.5.4 Antimicrobial activity*

AMPs including Beta-defensins have been studied to find out their antimicrobial activity and potency against different microbial organisms. Several differing assays have been used to investigate this. Table 1.1 summarises a selection of the research done in this area.



Organism and AMP	Organism(s) targeted	GRAM +/-	Minimum Inhibitory Concentration (MIC)	Assay used	Reference
Chicken Heterophil peptide 1 HCP1	<i>S. aureus</i> <i>E. coli</i>	+ -	5.3µg/ml 5.3µg/ml	Colourimetric assay 620nm	Evans <i>et al.</i> 1995
Crab-eating Macaque <i>Macaca fascicularis</i> β- defensin mfa	<i>E. coli</i> <i>P. aeruginosa</i> <i>S. aureus</i> <i>Candida albicans</i>	- - + n/a	4.0µM 4.0µM 8-16µM 16µM	Microdilution Susceptibility Test	Crovella <i>et al.</i> 2005
Turtle Egg White Protein (TEWP)	<i>E. coli</i> <i>S. typhimurium</i> <i>S. aureus</i>	- - +	3.3µM 2.8µM 5.1µM	IC <sub>50</sub>	Chattopadhyay <i>et al.</i> 2006

Avian Beta-defensin AvBD1	<i>B. subtilis</i> <i>B. cereus</i> <i>S. aureus</i> <i>S. haemolyticus</i> <i>S. saprophyticus</i> <i>L. monocytogenes</i>  <i>S. Enteritidis</i> ATCC 13076 <i>S. Enteritidis</i> LA5 <i>S. Typhimurium</i> <i>E. cloacae</i> <i>K. pneumoniae</i> <i>E. coli</i> <i>P. aeruginosa</i>	+ + + + + +  - - - - - -	0.19 $\mu$ M 0.21 $\mu$ M 0.08 $\mu$ M 0.14 $\mu$ M 0.16 $\mu$ M 0.26 $\mu$ M  0.17 $\mu$ M 0.16 $\mu$ M 0.15 $\mu$ M 0.20 $\mu$ M 0.10 $\mu$ M 0.27 $\mu$ M 0.27 $\mu$ M	Radial Diffusion Assay	Dereche <i>et al.</i> 2009
Chinese Softshell Turtle Pelovaterin	<i>S. aureus</i> <i>P. aeruginosa</i>	+ -	42.5 $\mu$ g/ml 0.42 $\mu$ g/ml	EC <sub>50</sub>	Lakshminarayanan <i>et al.</i> 2008
European Pond Turtle	<i>E. coli</i> ML35p	-	0.65(a), >20(b) $\mu$ Mol/L	Radial Diffusion	C. Stegemann <i>et al.</i> 2009

TBD-1	<i>L. monocytogenes</i> EGD	+	0.65(a), >20(b) $\mu\text{Mol/L}$	Assay	
	MRSA ATCC 33591	+	5.6(a), >20(b) $\mu\text{Mol/L}$		
	<i>Candida albicans</i> 820	n/a	5.2(a), >20(b) $\mu\text{Mol/L}$		
Human Beta-defensin 2 HBD-2	<i>S. aureus</i> <i>B. subtilis</i> <i>E. coli</i> <i>Candida albicans</i>	+ + - n/a	All wells aliquoted with 10 $\mu\text{g}$ of HBD-2 and zone of inhibition measured. HBD-2 showed activity towards all organisms at pH7.5	Radial Diffusion Assay	Yount <i>et al.</i> 2009

**Table 1.1 Studies performed using Beta-defensin peptides showing activity levels and assays used.**

(a) Low salt - 10mmol/L sodium phosphate buffer (pH7.4)

(b) High salt - 10mmol/L sodium-phosphate buffer (pH7.4), 0.1mol/L sodium chloride.

#### 1.5.4 Immobilization of AMPs in surface coatings

In addition to the key uses outlined above, AMPs could also have a role to play in addressing problems that arise from medical devices and implants due to microbial growth, leading to biofilm formation and infection eradication of such problems has been a major cause for concern. Biofilms are very difficult to deal with and with the growing issue of biotic resistance the development of antimicrobial coating on such devices has gained a lot of attention in recent years. AMP's and their unique properties outlined above have gained particular attention, however, to fully exploit AMPs in the use of antimicrobial coatings on medical devices careful attention is needed to bring about successful utilisation in the medical setting. AMP immobilisation strategies recently explored will be discussed.

One such problem that must be taken into consideration is biofouling. Salwiczek *et al* (2014) describes that an important parameter in this is surface topography as this determines microbial attachment and the formation of biofilms and sets out four guidelines for using antimicrobial coatings which will have long term effectiveness. These are:

- 1) Provide a surface topography that is unfavourable for microbial attachment.
- 2) Prevent the adsorption of biomolecules.
- 3) Kill all microbes that manage to overcome this anti- adhesion barrier.
- 4) Not retain dead microbes on the material surface.

There are number of methods reported for the immobilisation of AMPs into coating of surfaces. In summary, most involve the use of three basic methods, 'Icing', 'Bottle brush' and 'layer-by-layer' (LBL). 'Icing' involves coupling the AMP to the surface to form a chemically stable coating (Bagheri *et al* 2009). This method offers an advantage of giving more stable attachment of the peptide to the surface substrate (Goddard *et al.* 2007). The 'Bottle brush' approach employs polymer resins such as polyethylene glycol (PEG) that bear reactive groups for the immobilisation of AMPs (Bagheri *et al* 2009). PEG is often used because of its anti-adhesiveness for a bacterial colony to stick to the surface (Kingshott *et al.* 2009). Flexible space within the PEG can facilitate free movement of the bound AMP and therefore can possibly enhance the AMP-bacterial interactions. 'LBL' is the process of sandwiching AMPs

between two polyionic polymers and this also allows the number of layers to be tailored accordingly (Etienne *et al* 2004) allowing a desired loading on the surface. An advantage of this method allows slow release of the peptide; however, a drawback is that the AMPs closest to the solid support may not be in contact with the given target, thus reducing activity (Perez Espitia *et al* 2012).

#### 1.5.5 Potential and current uses of AMPs in medical and food industries

The properties of AMPs and coating outlined above make them good candidates for investigation within the medical and food industries. Urinary catheters are used quite extensively in medicine, but they can cause several problems for patients who have them administered due to urinary tract infections (UTIs). To prevent this from happening catheters need to be replaced at regular intervals placing a burden on the healthcare system and raising costs associated with this. Current problems with antibiotic resistance have warranted a need to investigate what new strategies can be employed to decrease the aforementioned issues. AMPs are peptides and can be seen as not too dissimilar to enzymes. One such enzyme that has currently been explored is that of cellobiose dehydrogenase (CDH) (Thallinger *et al.* 2014). This particular enzyme can produce hydrogen peroxide and when immobilised on a silicon surface inhibited the growth of both *E. coli* and *S. aureus*. They also explored putting this enzyme into the lubricant which showed similar results. This would make administration for the prevention of infection an easy strategy. However, functionalising dressings is still a new area and should be investigated further.

Another route that could be explored is the use of AMPs in functionalised cotton gauzes. One notable study developed a new strategy for incorporating AMPs into polyelectrolyte multilayer films over cotton gauzes (Gomez *et al* 2015). This was achieved by using the layer-by-layer approach as described above using an alternate deposition of a polycation of chitosan and a polyanion of alginic acid sodium salt with various AMPs sandwiched in between. The result of this study found that the AMPs incorporated into the gauzes had a higher antimicrobial effect compared to the controls and furthermore were non-cytotoxic to normal dermal fibroblasts at the concentrations tested.

The food industry, food safety is of a growing concern and a large number of food borne pathogens cause hospitalisations and deaths each year. In recent years increased consumer

knowledge has caused an increase in demand for minimally processed food with natural additives due to several growing controversies arising from processed foods with chemical preservatives (Perez Espitia *et al* 2012). In food preservation, peptides could overcome these issues and have already been shown to be able to be incorporated in to packaging materials (Appendini *et al* 2002).

Functionalising surfaces and materials with AMPs should be managed with a carefully thought out development process, not only to evaluate the action but also their stability, releasing abilities and tuneable performance (Felgueiras *et al* 2017). Using bioactive components on these have several advantages over cheaper methods such as they are natural, non-toxic, non-reactive to living organisms (Singha *et al* 2017), but one downside to current coatings is that they are expensive to produce. In summary more research should be done in this area.

In summary Beta-defensins are part of the innate immune system and have been shown to bridge the gap between the innate immune system and the adaptive immune system. They have evolved over a very long period. They have a conserved structure which shows similarities across the phylogenetic tree and has shown promise in helping wound healing. This has demonstrated that future perspectives in this area of research could allow new AMP's to be developed so that they could help activate the innate immune system in a variety of treatments as we are currently on our last line of antibiotic treatments available due to antibiotic resistance which has become one of the most important crises to hit medicine in the modern age. With the advent of high-throughput genomics, new reptilian genomes have been sequenced and their transcriptomes determined. Consequently, the challenges faced due to antibiotic resistance bacteria, new and useful molecules that could be found in reptiles may hold the key to develop future antibiotics and antimicrobial materials (van Hoek, 2014).

## **1.6 Reptiles – An interesting, untapped resource.**

Reptiles are evolutionarily ancient and are found in many different, diverse, and challenging environments. Over time the immune system of reptiles has evolved independently. Reptiles lack lymph nodes and do not form germinal centres (Rios *et al*, 2015). Thus, in response to infection they rely much more heavily on their innate immune responses.

Reptiles tend to have a robust innate immune system which begins with a keratinised outer epidermis preventing the entry of microbes into the body. If this protective layer is breached, their innate immune response is triggered. This constitutes the first line of defence. However, their adaptive immune response is more simplistic of their mammalian counterparts but regardless of this, reptiles are remarkably resilient, producing a wide range of cells and molecules, including antimicrobial peptides. However, studies investigating innate immune sensing and downstream responses have predominantly focused on human or mammal models so further information on the mechanisms in reptiles could provide information into the evolution of the innate immune system. It has also long been known that to escape predation, lizards lose their tails and then regenerate a new tail from the wound site. Studies in the Anole Lizard have shown that beta-defensins are expressed in the wound tissue with minimal inflammation and appear to play a large role in the regeneration and wound healing (Alibardi *et al* 2012). Little research has been done on these animals and it is becoming clear that more work is needed by scientists to identify new and useful AMPs. Although, gene knock-out studies have been done in mice (Chromek *et al* 2012) there is no direct data in reptiles to demonstrate their importance in overall resistance to infection and survival. Despite Lizards being the largest group of reptiles, which possess great resistance to infection from their environments there is little information regarding their AMPs (Alibardi, 2010). Hence there is a gap in our knowledge that reptiles could fill and therefore there is a need to explore this untapped resource.

## RESULTS



**Chapter 2. Method development for Data mining and gene annotations of genomic DNA to establish Beta-defensin clusters.**

Using individual known genes, in a BLAST search approach, is long and laborious when searching for beta-defensins in a given genome. Also using this approach to search through EST databases is limiting as it does not give you genomic organisation of whether the matches are clustered together or not. By their very nature of being small and quite variable in sequence, beta-defensins can be very difficult to find within genomes where the cluster has not yet been characterised; therefore, a different approach had to be used. It was found in this work that the use of a concatemer approach to searching is a more desirable way of identifying potential beta-defensin clusters within any given genome.

Producing a concatemer of known beta-defensins provided an advantage over searching using individual genes. It was found that using a concatemer provided a greater chance of identifying a potential beta-defensin in a genome search because the template has greater length and diversity of that of a singular gene allowing the probability to increase. It also gave a likeliness that the process would identify the cluster within the genome as there would be several matches within the same region of DNA being investigated.

Due to the work characterising the chicken beta-defensins (Lynn *et al.* 2004) and with birds being the closest living relative to reptiles, the use of the chicken beta-defensins at the start any search would be beneficial in identifying beta-defensins in reptiles.

### **2.1 Initial Searches**

Chicken beta-defensin predicted peptide sequences were obtained from The European Bioinformatics Institute (EMBL-EBI) pfam database (<http://pfam.xfam.org/>) by searching for 'beta-defensins' in the key word search and selecting the top hit, from the list. Also, the Green Anole Lizard (*Anolis carolinensis*) beta-defensin predicted peptide sequences were obtained from the NCBI nucleotide database by searching for 'beta-defensins' and taking the protein sequence from the list. Both sets of sequences were placed into a concatemer as shown in Fig. 2.1. These were then used as the query sequences for the BLAST searches using the

tBLASTn program. This was opposed to the singular approach adopted by Xiao (2004) and Tang (2012). It was proposed that using a concatemer would ‘pull out’ a greater number of matches when searching through a given genome.



**Figure 2.1: Chicken Beta-defensin concatemer used as template for initial searches.**

*A single beta-defensin is highlighted. Other beta-defensins can be seen with the initiating methionine in the amino acid sequence.*

The first genome that was explored using this search strategy was the *Chrysemys picta bellii* (Western Painted Turtle) genome assembly Chrysemys\_picta\_bellii-3.0.3 (RefSeq assembly Accession GCA\_000241765.3) submitted by the Painted Turtle Genome Sequencing Consortium on 31/03/2014. Making use of the tBLASTn function and employing the concatemer as the query sequence as opposed to using single genes yielded more matches within the search.

Figure 2.2 displays the result of the matches that are obtained from using a single beta-defensin in the BLAST search. Small regions of similarity were identified, but did not resemble beta-defensins, which contain the typical conserved cysteine motif. Utilisation of the concatemer increased the likelihood that more matches would appear during the search and within one scaffold in the assembly, narrowing down the identification of the clustered beta-defensins (figure 2.3).

Score	Expect	Method	Identities	Positives	Gaps	Frame
32.0 bits(71)	0.89	Compositional matrix adjust.	14/23(61%)	16/23(69%)	0/23(0%)	+2
Query 45	WPYYRVGSCGSLKSCCVRNRWA		67			
	WP+ R+G GS SCCVR RWA					
Sbjct 49733	WPWLRRLGRGGSPLSCCVRPRWA		49801			

Score	Expect	Method	Identities	Positives	Gaps	Frame
29.6 bits(65)	6.5	Compositional matrix adjust.	13/37(35%)	17/37(45%)	0/37(0%)	+3
Query 24	FSSPIHACRYQRGVCIPGPCRWPYYRVGSCGSLKSC		60			
	+ P+ AC RG GP RW + R G L +C					
Sbjct 693150	LAEPMQACTPHRGAATSGPGRWQVVRWGVVMGELGAC		693260			

**Figure 2.2. BLAST search results from single Beta-defensin used in query sequence.**

The results of the BLAST search using the concatemer yielded matches on 3 different scaffolds, with the scaffold in figure 2.3 being the most likely to harbour a Beta-defensin cluster. This is for two reasons, the first being, that the matches show the core cysteine motif that beta-defensins contain (indicated by arrows) and that there are several similar matches along a large stretch of DNA on the scaffold. Sorting by their starting position on the scaffold shows that the first match is at position 3135174 and the last match is at 3783001 covering a region of roughly 647kbps within the 4030082 base pairs of the identified scaffold.

Download ▾ GenBank Graphics Sort by: Subject start position ▾

Chrysemys picta bellii isolate RCT428 unplaced genomic scaffold, Chrysemys\_picta\_bellii-3.0.3  
Sequence ID: [NW\\_007281425.1](#) Length: 4030082 Number of Matches: 5

Range 1: 3135037 to 3135174 [GenBank](#) [Graphics](#) ▾ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
56.6 bits(135)	8e-06	Compositional matrix adjust.	26/46(57%)	31/46(67%)	0/46(0%)	-3
Query	521	APARGFSDSQLCRNNHGHCRRLCFHMESWAGSCMNGRLRCCRFSTK	566			
		A G D+ C +NHGHCRRLCFHME G+C NG LRCC+ T+				
Sbjct	3135174	AARGGTDTLQCLSNHGHCRRLCFHMEHQVGTCTNGHLRCC*ETR	3135037			
		↑ ↑ ↑ ↑ ↑				

Range 2: 3435575 to 3435715 [GenBank](#) [Graphics](#) ▾ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
39.3 bits(90)	1.1	Compositional matrix adjust.	19/49(39%)	27/49(55%)	2/49(4%)	-2
Query	742	PGSADPLFPDTPVACRTQGNFCRAGACPPTFTISGQCHGGLLNCCAKIPA	790			
		G+AD F D + CR+ FC +G CP + T+ G C G +NCC +				
Sbjct	3435715	SGNAD--FLDNINCRSNFGFCHSGDCPISTTLIGTCINGKINCKRRTT	3435575			
		↑ ↑ ↑ ↑ ↑				

Range 3: 3447223 to 3447363 [GenBank](#) [Graphics](#) ▾ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
52.8 bits(125)	9e-05	Compositional matrix adjust.	26/47(55%)	31/47(65%)	2/47(4%)	+1
Query	743	GSAD--PLFPDTPVACRTQGNFCRAGACPPTFTISGQCHGGLLNCCAK	787			
		G AD P DT+AC+ QG FCR CPP F++SG CHGG L CC +				
Sbjct	3447223	GVADVGPPPADTLACKAQGGFCRLLNCPVFSVSGTCHGGQLQCCTR	3447363			
		↑ ↑ ↑ ↑ ↑				

Range 4: 3461757 to 3461879 [GenBank](#) [Graphics](#) ▾ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
36.6 bits(83)	7.3	Compositional matrix adjust.	20/42(48%)	23/42(54%)	1/42(2%)	+3
Query	676	PGLSLARGLPQDCERRGGFCSHKSCPPGIGRIGLCSKEDFCC	717			
		G + A+ + Q C R GG C SCP G RIG CS D CC				
Sbjct	3461757	AGFTQAQNIQ-CIRLGGSCRSGSCPSGFARIGTCSGSDSCC	3461879			
		↑ ↑ ↑ ↑ ↑				

Range 5: 3782861 to 3783001 [GenBank](#) [Graphics](#) ▾ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
37.4 bits(85)	5.3	Compositional matrix adjust.	19/47(40%)	21/47(44%)	0/47(0%)	-2
Query	677	GLSLARGLPQDCERRGGFCSHKSCPPGIGRIGLCSKEDFCCRSRWYS	723			
		G + P C R GGFC CPP RIG C CC+ W S				
Sbjct	3783001	GFTQGINTPFACRRAGGFCRRGRCPNFRIRGSCGFGQSCCKRGWVS	3782861			
		↑ ↑ ↑ ↑ ↑				

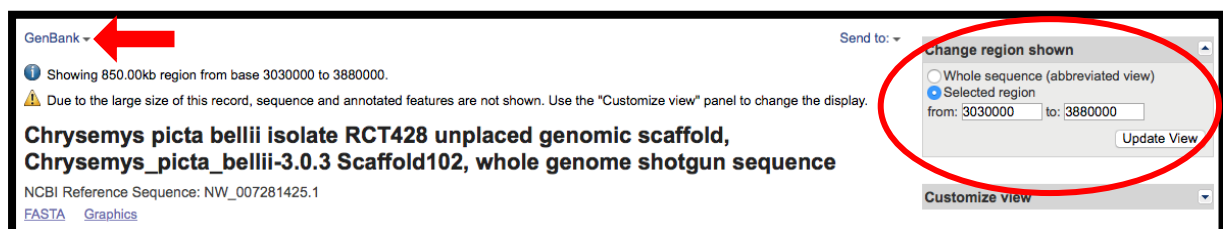
**Figure 2.3. BLAST search result from using Chicken concatemer.**

BLAST output sorted by start position showing 5 potential matches on a singular scaffold.

Blue arrows highlight Beta-defensin conserved cysteine motif.

## 2.2 First Annotations

At this point, the size of the cluster within this scaffold could start to be determined. Clicking 'sequence ID' on the results page, in this case NW\_007281425.1, opens a landing page for the scaffold information. Here it was possible to select a region of the scaffold for further analysis. Using the positions of the matches shown circled in figure 2.3 a region of 100 kb upstream of the first match and 100 kb downstream of the last match was then selected. This gave a 'buffer' to allow for any beta-defensins that the first BLAST search may have missed. This resulted in a region of around 850 kb in which more iterative searching could be done. Figure 2.4 shows the page in which this was achieved.



**Figure 2.4. Page showing how region on scaffold is changed.**

Selection of 'FASTA (text)' from the drop-down menu (marked by arrow in Figure 2.4) allowed the sequence to be viewed. This was then copied into a new iterative BLAST search, once again using the chicken beta-defensin concatemer as a query. By narrowing down the search from the whole genome to the size of the potential cluster, the program could bring out significantly more matches, rising from 5 in the first genomic search to 76 in the search of this 850kbp region. It also identified potential first exon signal peptides of the beta-defensin gene (Figure 2.5A). The first identified match (Figure 2.5B) showed that the position of 105038 kb from the start of the DNA sequence and the last match identified (Figure 2.5C) by this search indicated a position of 794230 kb from the start of the DNA sequence. This gave an estimated size of the cluster at 700 kb on length. This DNA sequence was reduced further by decreasing the size. A 'buffer' of 50 kb was kept at each end of the DNA sequence to allow for any potential matches. The range selected on this scaffold was from positions 3080000-3830000 for the next, more in-depth, searches.

A)

Range 4: 108145 to 108225 <a href="#">Graphics</a>		▼ Next Match ▲ Previous Match ▲ First Match				
Score	Expect	Method	Identities	Positives	Gaps	Frame
26.9 bits(58)	2.3	Compositional matrix adjust.	13/30(43%)	18/30(60%)	3/30(10%)	-3
Query	993	MRIVYLLIPFFLLFLQGAAGTATQCRIRGG	1022			
		MRI+YL + FLQ A+G + +I GG				
Sbjct	108225	MRILYLFFAVVIFFLQAASG---EVKISGG	108145			

B)

Range 1: 105038 to 105175 <a href="#">Graphics</a>		▼ Next Match ▲ Previous Match				
Score	Expect	Method	Identities	Positives	Gaps	Frame
56.6 bits(135)	3e-09	Compositional matrix adjust.	26/46(57%)	31/46(67%)	0/46(0%)	-2
Query	521	APARGFSDSQLCRNNHGHCRRLCFHESWAGSCMNGRLRCCRFSTK	566			
		A G D+ C +NHGHCRRLCFHME G+C NG LRCC+ T+				
Sbjct	105175	AARGGYDTLQCLSNHGHCRRLCFHMEHQVGTCTNGHLRCC*ETR	105038			

C)

Range 76: 794230 to 794376 <a href="#">Graphics</a>		▼ Next Match ▲ Previous Match ▲ First Match				
Score	Expect	Method	Identities	Positives	Gaps	Frame
25.8 bits(55)	5.1	Compositional matrix adjust.	17/50(34%)	22/50(44%)	1/50(2%)	-3
Query	513	VVILLLQDAPARGFSDSQLCRNNHGHCRRLCFHESWAGSCMNGRLRCCR	562			
		+ L L + S+ CR G C R+CF G+C G L CCR				
Sbjct	794376	IKTLFLCAGTFEFINSSRACRRARGSCFRVCFRRYRLIGTCQG-LSCCR	794230			

**Figure 2.5. Results from BLAST search from smaller 850kbp stretch of Scaffold**

Using Chicken Beta-defensin concatemer. A) Potential first exon signal peptide match. B) Initial second exon match and position. C) Last match and position.

At this stage of initial searches and determining a potential stretch of DNA containing the cluster, the next stage of the identification process could begin. This involved the generation of a six-frame translation of the DNA sequence of interest. This was done as the program will convert the DNA sequence into a translation of all reading frames allowing annotations to be highlighted and populated. To generate the six-frame translation of the DNA sequence the FASTA file (highlighted by arrow in figure 2.4) was copied into STEP 1 box of the program landing page (figure 2.6). The Program EMBOSS Sixpack (Madeira *et al.* 2019) was utilised for this and can be found at The European Bioinformatics Institute (EMBL-EBI) website under 'find a tool'. [https://www.ebi.ac.uk/Tools/st/emboss\\_sixpack/](https://www.ebi.ac.uk/Tools/st/emboss_sixpack/)



Tools > Sequence Translation > EMBOSS Sixpack

## EMBOSS Sixpack

Sixpack reads a DNA sequence and outputs the three forward and (optionally) three reverse translations in a visual manner.

STEP 1 - Enter your input sequence

Enter or paste a  sequence in any supported format:

Or, [upload a file](#):  no file selected

[Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

**Figure 2.6. Input box for FASTA file sequence in EMBOSS Sixpack program website.**

Score	Expect	Method	Identities	Positives	Gaps	Frame
56.6 bits(135)	2e-09	Compositional matrix adjust.	26/46(57%)	31/46(67%)	0/46(0%)	-2
Query 521	APARGFSDSQLCRNNHGHCRRLCFHMESWAGSCMNGRLRCCRFSTK		566			
	A G D+ C +NEGHCRRLCFHME G+C NG LRCC+ T+					
Sbjct 55175	AARGGTYDTLQCLSNHGHCRRLCFHMEHQVGTCTNGHLRCK*ETR		55038			

Score	Expect	Method	Identities	Positives	Gaps	Frame
26.9 bits(58)	2.0	Compositional matrix adjust.	13/30(43%)	18/30(60%)	3/30(10%)	-3
Query 993	MRIVYLLIPFFLLFLQGAAGTATQCRIRGG		1022			
	MRI+YL + FLQ A+G + +I GG					
Sbjct 58225	MRILYLFFAVVIFFLQAASG---EVKISGG		58145			

**Figure 2.7. Positions of potential first and second exon match.**

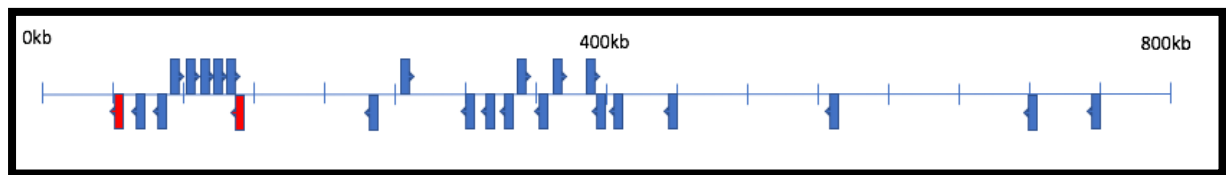
Taken from region to be annotated onto 6-frame translation from EMBOSS. These matches are on the reverse strand showing the second exon match coming before the first exon match.

Once processed an output file could be downloaded. This output file was then annotated with the matches from the BLAST searches using positions shown. As described previously the chicken concatemer was used as a query sequence using BLAST then followed by an iterative search using a concatemer of the Green Anole Lizard defensin peptide sequences. Using the output file one was able to begin to highlight the matches onto the reading frame within it (figure 2.8) with the search matches identified by the BLAST from the reduced size region of the scaffold (positions 3080000-3830000) as shown in figure 2.4.





both 1<sup>st</sup> and 2<sup>nd</sup> exon on the structure. Table 2.1. shows locations of potential matches along the DNA scaffold.



**Figure 2.9. Genomic Map of BLAST matches**

*Plot of BLAST matches from both Chicken and Anole Lizard Concatemers. Red arrows show where both potential 1<sup>st</sup> and 2<sup>nd</sup> exon for a particular beta-defensin gene are found. Blue arrows show only potential 2<sup>nd</sup> exon matches. Above and below line show transcript direction with direction arrow.*

First/Signal approx position	2nd Exon approx position	Direction
58145	55038	-
	67163	-
	84782	-
	93118	+
	104400	+
	112768	+
	118800	+
	125528	+
141891	139736	-
	235297	-
	265565	+
	310254	-
	320053	-
	329451	-
	341113	+
	355576	-
	367224	+
	381698	+
	396079	-
	411941	-
	442634	-
	533630	-
	564874	-
	702868	-
	744230	-

**Table 2.1. Positions of potential matches against tBLASTn searches using the Chicken and Green Anole Lizard.**

### **2.3.1 GENSCAN**

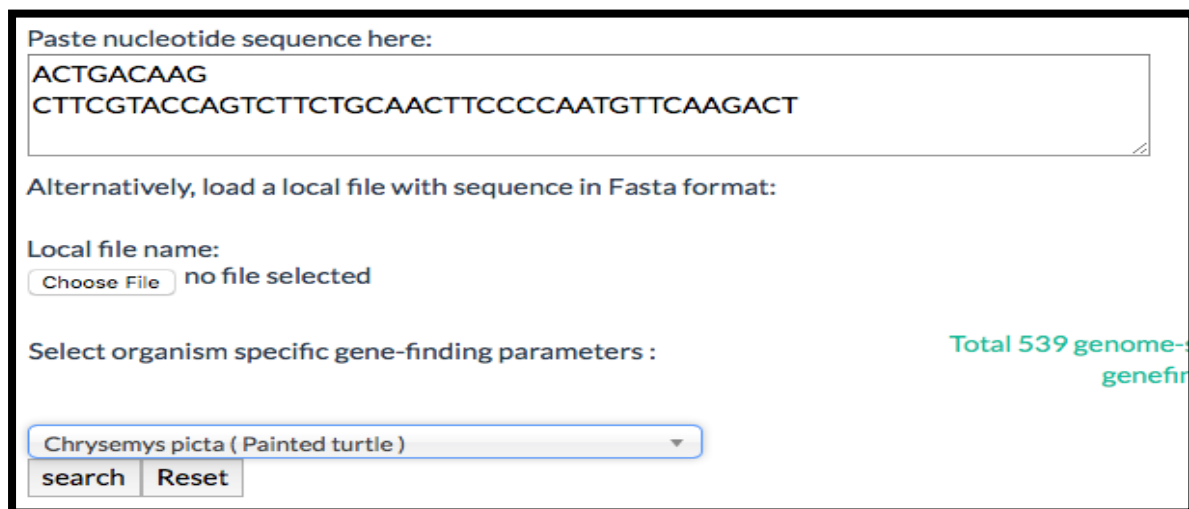
The GENSCAN server at MIT was used for the identification of complete gene structures in genomic DNA. The program GENSCAN predicts the locations and exon-intron structures of genes in genomic sequences from a variety of organisms. This program was used to further analyse the sequences. The GENSCAN results were annotated onto the EMBOSS file with all the possible exons identified so far. The data showed that there were still exons to be identified further. This led to the use of the FGENESH program (Solovyev *et al.* 2006).

### 2.3.2 Softberry – FGENESH

Softberry, Inc. is a leading developer of software tools for genomic research. FGENESH uses HMM-based gene structure prediction in a similar way to GENSCAN with a similar output. From the homepage, copy and paste the sequence to be analysed and select what organism to use for the gene finding parameters and click 'search' (figure 2.10).

The output gave a list of positions of the predicted genes and further down the page gave a list of predicted proteins with their nucleotide sequences. Looking at the FGENESH prediction results and comparing to gene identified using GenScan it was possible to see that some of the potential gaps were filled. Figure 2.11 shows one such example.

FGENESH identifies the gene positions without having to BLAST against the original scaffold if it predicts a known gene. However, it may be necessary to follow a similar principle as the GENSCAN technique to further find potential matches. Table 1.2 and figure 2.12 shows the final table of positions on the scaffold and the final genomic map of identified genes.



Paste nucleotide sequence here:  
ACTGACAAG  
CTTCGTACCAGTCTTCTGCAACTTCCCAATGTTCAAGACT

Alternatively, load a local file with sequence in Fasta format:  
Local file name:  
Choose File no file selected

Select organism specific gene-finding parameters : Total 539 genome-wide genes identified

Chrysemys picta ( Painted turtle )

search Reset

**Figure 2.10** Input field for FGENESH.

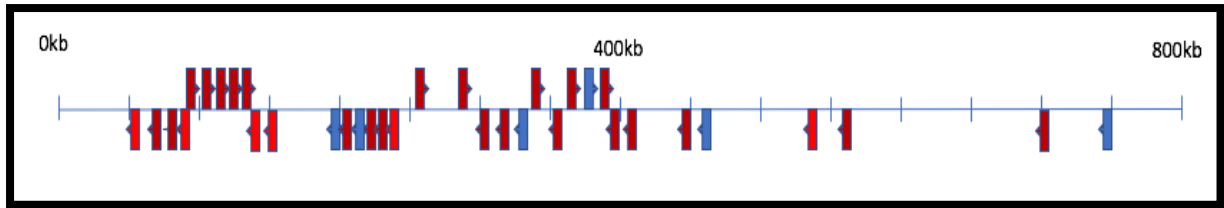
*Nucleotide sequence is copied and species selected for gene identification.*

```
>FGENESH:[mRNA] 7 2 exon (s) 67127 - 69096 189 bp, chain
ATGAAGATCTTTATCTTCTTTTGCTGGTCTCTTCTTGGTCTCCCTGCCCAATCCAGGG
AATGGCCAGTTTGTCAATTCTGGGCTGCTTAATCAGAGGCGGTTCTGTCGTAAGCAAT
TGCTACCTGGATGAAACGGAGATCGGGAGCTGCCTAAGGAGCAATAGGAAATGCTGCAAG
ACAACATGA
>FGENESH: 7 2 exon (s) 67127 - 69096 62 aa, chain -
MKIFYLLFAGLFLVSLPNPNGQFVILGCLIRGGSCRTDNCYLDETEIGSCLRSNRKCK
TT
```

**Figure 2.11.** Identification of Potential Beta-defensin gene missing from GenScan analysis.

First/Signal approx. Position	2 <sup>nd</sup> Exon approx. Position	Direction
58145	55038	-
69096	67163	-
75698	74930	-
85724	84782	-
88608	93118	+
102451	104400	+
111636	112768	+
114239	118800	+
124484	125528	+
141891	139736	-
152993	151850	-
197342		-
206797	202725	-
217748		-
220760	219159	-
224974	223326	-
236266	235297	-
264638	265565	+
282578	285545	+
311665	310254	-
320413	320053	-
	329451	-
339957	341113	+
357225	355576	-
364871	367224	+
371876		+
379335	381698	+
397726	396079	-
414544	411941	-
444566	442634	-
463948		-
537049	533630	-
569278	564874	-
706278	702868	-
	744230	-

**Table 2.2. Full list of matches/predictions from BLAST searching and Gene Prediction software.**



**Figure 2.12. Full genomic map of Predicted Genes.**

## **2.4 Splice Site Prediction**

Once the gene finding was exhausted, the next stage in the identification and annotation process was to predict the splice sites of the predicted genes. Splice site prediction in this analysis was done using the online server by the Berkeley Drosophila Genome Project - [http://www.fruitfly.org/seq\\_tools/splice.html](http://www.fruitfly.org/seq_tools/splice.html) (Reese *et al.* 1997) and Softberry FSPLICE - <http://www.softberry.com/berry.phtml?topic=fsplce&group=programs&subgroup=gfind>.

The genomic sequence starting from the initiating methionine to roughly 60 bp past the identified match from the blast on the 1<sup>st</sup> exon was copied and pasted into the input box and the 60 bp upstream from the start of the 2<sup>nd</sup> exon identified to the stop codon was copied into the input box on the web page (figure 2.13). The example used here was the first match from the previous searches. As this example was on the reverse strand the raw sequence for the analysis will be taken from the forward strand, so when submitting sequences to the server, check the box 'include reverse strand'. As an alternative, the complementary strand could be submitted; this method was used as the example.

The output file figure 2.14 shows several predicted sequences taken from the input and also highlights the splice sites. These can then be annotated, once again, on the EMBOSS output file. What is known about Beta-defensin structure is that the 1<sup>st</sup> exons are relatively short so the first in the predicted donor splice sites seem fits this rationale as it was only around 58 bp long. Also, the score of the prediction comes out a 0.98 (figure 2.14).

Looking at the output of the acceptor site predictions there were more potential options. It was known that the donor site is between positions 52-66 so it was likely that the first 3 predictions can be discarded as this region was taken from the same stretch of DNA as the Donor Site.

**Organism:**  Drosophila  Human or other  
**Search for 5', 3' or both splice sites?**  3'  5'  both  
**Include reverse strand?**  yes  no  
**Minimum score for 5' splice site (between 0 and 1):**   
**Minimum score for 3' splice site (between 0 and 1):**

**Cut and paste your sequence(s) here:** Use single-letter nucleotides: (A, C, G, T).  
 You can include multiple sequences if each has a FASTA title line starting with >

```

ATGAGGATCCTTTACCTGTTCTTTGCTGTTGTCATCTTCTTCCCTCCAG
GCTGCTTCAGGTGAGGTGAAAATATCTGGAGGGCTTTTATAGGATGT
ATGGAGAGGGGAAAGAAATGTTGGGGCTCTGCAGGGCTCCACCCT
TCTCACTGCAACTATCTCACGTTTCCTTTCCTTGATAGCAAGAGG
CGGCACTTACGACACCTTGCACTGCTGAGCAACCATGGCCACTGC
CGACGGCTTTGCTTCCACATGGAACATCAGGTTGGCACCTGCACCA
ATGGTCACCTGCGCTGCTGCAAG|
  
```

**Figure 2.13** Splice site prediction input box with DNA sequence.

**Donor site predictions for 109.180.3.178.31785.0 :**

Start	End	Score	Exon	Intron
52	66	0.98	gcttcag	gtgaggtg
257	271	0.47	acatcag	gttggcac

---

**Acceptor site predictions for 109.180.3.178.31785.0 :**

Start	End	Score	Intron	Exon
28	68	0.96	gttgatcattcttctctcc	aggctgcttcaggtgaggtgaa
38	78	0.61	tcttctctccaggctgcttc	agggtgaggtgaaaatctgga
70	110	0.43	atatctggagggtctttat	aggatgtatggagaggggaaag
109	149	0.61	agaaatgttggggctctgc	agggtccacccttctcactgc
160	200	0.96	cgtttctttcttctgtagc	agcaagagggcggcacttacgac

**Figure 2.14.** Splice site prediction output file.

Predictions from inputted DNA sequences. Both predicted donor and acceptor sites, along with scores shown.

```

K A L Q I F S P H L K Q P G G R R * Q Q F1
K P S R Y F H L T * | S S L E E E D D N S F2
S P P D I F T S P E A A W R K K M T T A F3
58141 AAAGCCCTCCAGATATTTTCACCTCACCTGAAGCAGCCTGGAGGAAGAAGATGACAACAG 58200
-----:-----|-----:-----|-----:-----|-----:-----|-----:-----|-----:-----|
58141 TTTCGGGAGGTCTATAAAAGTGGAGTGGACTTCGTCGGACCTCCTTCTTCTACTGTTGTC 58200
F A R W I N E G * R F C G P P L L H C C F6
L L G G S I K V E G S A A Q L F F I V V F5
F G E L Y K * R V Q L L R S S S S S L L F4

Q R T G K G S S W V N V S L G S R S H T F1
K E Q V K D P H G * M S A W D L G V T P F2
K N R * R I L M G E C Q L G I S E S H Q F3
58201 CAAAGAACAGGTAAAGGATCCTCATGGGTGAATGTCAGCTTGGGATCTCGGAGTCACACC 58260
-----:-----|-----:-----|-----:-----|-----:-----|-----:-----|-----:-----|
58201 GTTTCTTGTCCATTTTCTAGGAGTACCCACTTACAGTCGAACCCTAGAGCCTCAGTGTGG 58260
C L V P L P D E H T F T L K P D R L * V F6
A F F L Y L I R M P S H * S P I E S D C F5
L S C T F S G * P H I D A Q S R P T V G F4

```

Figure 2.15. Donor Splice site annotated on EMBOSS file.

```

S V L H P S L V S H L Q Q R R * P L V Q F1
A F F I P A L S L T C S S A G D H W C R F2
R S S S Q P C L S L A A A Q V T I G A G F3
55021 AGCGTTCTTCATCCCAGCCTTGTCTCTCACTTGCAGCAGCGGATGACCATTGGTGCAG 55080
-----:-----|-----:-----|-----:-----|-----:-----|-----:-----|
55021 TCGCAAGAAGTAGGGTCGGAACAGAGAGTGAACGTCGTCGCGTCCACTGGTAACCACGTC 55080
L T R * G L R T E * K C C R L H G N T C F6
S R E E D W G Q R E S A A A C T V M P A F5
A N K M G A K D R V Q L L A P S W Q H L F4

V P T * C S M W K Q S R R Q W P W L L R F1
C Q P D V P C G S K A V G S G H G C S G F2
A N L M F H V E A K P S A V A M V A Q A F3
55081 GTGCCAACCTGATGTTCCATGTGGAAGCAAAGCCGTCGGCAGTGGCCATGGTTGCTCAGG 55140
-----:-----|-----:-----|-----:-----|-----:-----|-----:-----|
55081 CACGGTTGGACTACAAGGTACACCTTCGTTTCGGCAGCCGTCACCGGTACCAACGAGTCC 55140
T G V Q H E M H F C L R R C H G H N S L F6
P A L R I N W T S A F G D A T A M T A * F5
H W G S T G H P L L A T P L P W P Q E P F4

H C K V S * V P P L A A T R K G N V R * F1
T A R C R K C R L L L L Q G K E T * D S F2
L Q G V V S A A S C C Y K E R K R E I V F3
55141 CACTGCAAGGTGTCGTAAGTGCCGCCTCTTGCTGCTACAAGGAAAGGAAACGTGAGATAG 55200
-----:-----|-----:-----|-----:-----|-----:-----|-----:-----|
55141 GTGACGTTCCACAGCATTACGGCGGAGAACGACGATGTTCCCTTTCCTTGCCTCTATC 55200
C Q L T D Y T G G R A A V L F P F T L Y F6
A S C P T T L A A E Q Q * L S L F R S I F5
V A L H R L H R R K S S C P F S V H S L F4

```

Figure 2.16. Acceptor splice site annotated on EMBOSS output.

Interpretation of the next two predictions came down to the structure of beta-defensins in the fact that there is a common motif of 6 cysteines within the 2<sup>nd</sup> exon. Working back from the end of the exon and counting back it shows that the splice site is likely to be the last one marked. This is also supported by the score being 0.96. Thus, the potential final sequence of the predicted gene can be produced and translated.

#### ***DNA sequence***

```
ATGAGGATCCTTTACCTGTTCTTTGCTGTTGTCATCTTCTTCCTCCAGGCTGCTTCAGCAGCAAGAGG
CGGCACTTACGACACCTTGCAGTGCCTGAGCAACCATGGCCACTGCCGACGGCTTTGCTTCCACATG
GAACATCAGGTTGGCACCTGCACCAATGGTCACCTGCGCTGCTGCAAG
```

#### ***Translation***

```
MRILYLFFAVVIFFLQAASAARGGTYDTLQCLSNHGHCRRLCFHMEHQVGTCTNGHLRCK
```

Finally, search beyond the stop codon at the end of the exon for the poly adenylation signal. This will ensure that you have not missed out any other exons that may possibly be there as some beta-defensins have been found to have a small 3<sup>rd</sup> exon at the end (figure 2.17).

So, from the genes that were predicted from the above scaffold these could then be put into a concatemer and then final iterative BLAST searches could be performed to see if any final missing exons could be found. Once a degree of confidence is established that all the genes within the cluster have been identified and to go further, a repeat masker was then performed to establish any larger gaps between repetitive sequences that may have the presence of beta-defensins. This would then ensure that the cluster has been annotated as fully as possible.



```

F I C N G K S M S L V S L V S S M G V T F1
L F A M E K A * V W C R W C P A W E L H F2
Y L Q W K K H E F G V A G V Q H G S Y T F3
54901 TTTATTTGCAATGGAAAAAGCATGAGTTTGGTGTGCTGGTCCAGCATGGGAGTTACA 54960
-----:-----|-----:-----|-----:-----|-----:-----|-----:-----|
54901 AAATAAACGTTACCTTTTTCTACTCAAACCACAGCGACCACAGGTCGTACCCTCAATGT 54960
K I Q L P F L M L K T D S T D L M P T V F6
N * K C H F F C S N P T A P T W C P L * F5
K N A I S F A H T Q H R Q H G A H S N C F4

R V R C A V I G K S C S V S F R L W V Q F1
V F A V L L L E S L V R F P S A F G S R F2
C S L C C Y W K V L F G F L P P L G P E F3
54961 CGTGTTGCTGTGCTGTTATTGGAAAGTCTTGTTCGGTTTCCTTCCGCCTTTGGGTCCAG 55020
-----:-----|-----:-----|-----:-----|-----:-----|-----:-----|
54961 GCACAAGCGACACGACAATAACCTTTTCAGAACAAGCCAAAGGAAGGCGGAAACCCAGGTC 55020
R T R Q A T I P F D Q E T E K R R Q T W F6
V H E S H Q * Q F T K N P K R G G K P G F5
T N A T S N N S L R T R N G E A K P D L F4

S V L H P S L V S H L Q Q R R * P L V Q F1
A F E I P A L S L T C S S A G D H W C R F2
R S S S Q P C L S L A A A Q V T I G A G F3
55021 AGCGTTCCTTCATCCCAGCCTTGTCTCTCACTTGCAGCAGCGCAGGTGACCATTGGTGCAG 55080
-----:-----|-----:-----|-----:-----|-----:-----|-----:-----|
55021 TCGCAAGAAGTAGGGTCGGAACAGAGAGTGAACGTCGCTCCACTGGTAACCACGTC 55080
L T R * G L R T E * K C C R L H G N T C F6
S R E E D W G Q R E S A A A C T V M P A F5
A N K M G A K D R V Q L L A P S W Q H L F4

```

**Figure 2.17. Poly adenylation signal highlighted on EMBOSS file.**

*In yellow the end of the second exon is shown.*

### 2.5 Repeat Masking

RepeatMasker (<http://www.repeatmasker.org/>) is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output shows the positions and which part of a repeat is present in an analysed DNA sequence. This was used as it highlighted the repeat sequences within the scaffold. Generally, repeat sequences are not in coding sequences so this was a way of masking out a proportion of the scaffold that did not have coding exons in it.

To use RepeatMasker on the web server go to <http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker>. It will bring up an input page.

From this page, upload the sequence of interest or copy and paste into the text box. This example uses the scaffold identified previously. Once 'rmbblast' and select 'vertebrate (other than blow)' was chosen, the sequence inputted was submitted for analysis.

Once the program had processed the sequence of interest an email with a web link was received, the link would lead to a results page where a summary (figure 2.18) and links to the output spreadsheet were given.

**Summary:**

```

=====
file name: RM2sequpload_1582276978
sequences: 1
total length: 800001 bp (578673 bp excl N/X-runs)
GC level: 45.55 %
bases masked: 100989 bp ( 12.62 %)
=====

```

	number of elements*	length occupied	percentage of sequence
Retroelements	231	83889 bp	10.49 %
SINEs:	46	6828 bp	0.85 %
Penelope	0	0 bp	0.00 %
LINEs:	178	76736 bp	9.59 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	178	76736 bp	9.59 %
R1/LOA/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	0	0 bp	0.00 %
RTE/Bov-B	0	0 bp	0.00 %
L1/CIN4	0	0 bp	0.00 %
LTR elements:	7	325 bp	0.04 %
BEL/Pao	0	0 bp	0.00 %
Tyl/Copia	0	0 bp	0.00 %
Gypsy/DIRS1	4	202 bp	0.03 %
Retroviral	3	123 bp	0.02 %
DNA transposons	79	9133 bp	1.14 %
hobo-Activator	44	5338 bp	0.67 %
Tcl-IS630-Pogo	5	450 bp	0.06 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	9	1456 bp	0.18 %
Other (Mirage, P-element, Transib)	0	0 bp	0.00 %
Rolling-circles	0	0 bp	0.00 %
Unclassified:	9	1959 bp	0.24 %
<b>Total interspersed repeats:</b>		<b>94981 bp</b>	<b>11.87 %</b>
Small RNA:	9	694 bp	0.09 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	103	4447 bp	0.56 %
Low complexity:	26	1136 bp	0.14 %

Figure 2.18. Summary of repeats identified from repeat masker.

By scrolling down to the bottom of the page, click on the text output where it gave a list of repeats and their positions on the scaffold (figure 2.19).

SW score	perc div.	perc del.	perc ins.	query sequence	position in query begin end	matching repeat (left)	repeat class/family	position in repeat begin end (left)	ID
627	7.4	0.0	0.0	NW_007281425.1:3080000-3880000	1094 1174	(798827) C UC0N81	DNA/hAT-Charlie	(0) 84	4 1
305	16.4	0.0	1.8	NW_007281425.1:3080000-3880000	4499 4554	(795447) C MIR	SINE/MIR	(111) 151	97 2
388	9.4	0.0	0.0	NW_007281425.1:3080000-3880000	4674 4726	(795275) C Plat_L3	LINE/CR1	(15) 3558	3506 3
1696	14.8	0.7	3.8	NW_007281425.1:3080000-3880000	5270 5570	(794431) + TurRetro4a	Retroposon	3 294	(0) 4
2453	3.8	0.0	0.0	NW_007281425.1:3080000-3880000	8653 8945	(791056) C TurRetrole	Retroposon	(0) 293	1 5
317	27.9	8.5	0.0	NW_007281425.1:3080000-3880000	9722 9850	(790151) C MIR_Testu	SINE/MIR	(16) 248	109 6
15	0.0	4.5	0.0	NW_007281425.1:3080000-3880000	12137 12158	(787843) + (GCC)n	Simple_repeat	1 23	(0) 7
12	9.4	8.3	2.6	NW_007281425.1:3080000-3880000	12486 12521	(787480) + (CCCAGCT)n	Simple_repeat	1 38	(0) 8
522	13.2	0.3	2.1	NW_007281425.1:3080000-3880000	15565 15657	(784344) C CR1-L3B_Croc	LINE/CR1	(1) 3276	3197 9
1036	15.8	0.5	2.0	NW_007281425.1:3080000-3880000	20486 20685	(779316) C MIR1_Amm	SINE/MIR	(31) 199	3 10
35	21.6	4.8	1.9	NW_007281425.1:3080000-3880000	21725 21829	(778172) + (TG)n	Simple_repeat	1 108	(0) 11
30	0.0	0.0	0.0	NW_007281425.1:3080000-3880000	21847 21872	(778129) + (TG)n	Simple_repeat	1 26	(0) 11
30	0.0	0.0	0.0	NW_007281425.1:3080000-3880000	21893 21918	(778083) + (TG)n	Simple_repeat	1 26	(0) 11
31	0.0	0.0	0.0	NW_007281425.1:3080000-3880000	22204 22230	(777771) + (GT)n	Simple_repeat	1 27	(0) 12

Figure 2.19. Output file showing positions of repeats on the scaffold of interest.

This was then copied to an excel spreadsheet. The reason behind this was to look closely into the gaps between the repeats so that it can be seen if any genes were missed during the initial searches. Once the information was copied over two columns were added. One to give the size of the repeat fragment and the other to give the size of the gap between repeat sequences. Once this was achieved the next step is to populate the spreadsheet with the exons of the predicted beta-defensins found during the search and annotation process. This would involve going back to NCBI BLAST and using the DNA sequence of the Beta-defensins identified to search against the scaffold used (figure 2.20)

Range 5: 55050 to 55176 <a href="#">Graphics</a>		▼ Next Match	▲ Previous Match	▲ First Match
Score	Expect	Identities	Gaps	Strand
244 bits(127)	3e-66	127/127(100%)	0/127(0%)	Plus/Minus
Query 57	AGCAGCAAGAGGCGGCACTTACGACACCTTGCACTGCCTGAGCAACCATGGCCACTGCCG			116
Sbjct 55176	AGCAGCAAGAGGCGGCACTTACGACACCTTGCACTGCCTGAGCAACCATGGCCACTGCCG			55117
Query 117	ACGGCTTTGCTTCCACATGGAACATCAGGTTGGCACCTGCACCAATGGTCACCTGCGCTG			176
Sbjct 55116	ACGGCTTTGCTTCCACATGGAACATCAGGTTGGCACCTGCACCAATGGTCACCTGCGCTG			55057
Query 177	CTGCAAG	183		
Sbjct 55056	CTGCAAG	55050		
Range 6: 58168 to 58225 <a href="#">Graphics</a>		▼ Next Match	▲ Previous Match	▲ First Match
Score	Expect	Identities	Gaps	Strand
112 bits(58)	2e-26	58/58(100%)	0/58(0%)	Plus/Minus
Query 1	ATGAGGATCCTTTACCTGTTCTTTGCTGTTGTCATCTTCTTCTCCAGGCTGCTTCAG			58
Sbjct 58225	ATGAGGATCCTTTACCTGTTCTTTGCTGTTGTCATCTTCTTCTCCAGGCTGCTTCAG			58168

**Figure 2.20. Positions of first Beta-defensin prediction.**

This was then inputted on the RepeatMasker spreadsheet input (figure 2.21) and repeated for the all the beta-defensins identified in this cluster region.

54	+	276	24.2	0	0	NW_007281425.1:3080000-3880000	55493	55558	-744443	65	34
55						<b>BD1 Exon 2</b>	<b>55050</b>	<b>55176</b>		<b>126</b>	<b>797</b>
56	+	302	18.6	1.6	3.3	NW_007281425.1:3080000-3880000	55592	55652	-744349	60	321
57	+	31	0	0	0	NW_007281425.1:3080000-3880000	55973	56001	-744000	28	353
58	+	256	24.6	0	1.4	NW_007281425.1:3080000-3880000	56354	56423	-743578	69	98
59	+	363	19.1	14.3	0	NW_007281425.1:3080000-3880000	56521	56604	-743397	83	117
60	+	2280	29.4	4.4	1.4	NW_007281425.1:3080000-3880000	56721	57658	-742343	937	1
61	+	397	14.9	1.3	1.3	NW_007281425.1:3080000-3880000	57659	57733	-742268	74	1557
62						<b>BD1 Exon 1</b>	<b>58168</b>	<b>58225</b>		<b>57</b>	<b>1731</b>
63	+	12	7.8	7.4	0	NW_007281425.1:3080000-3880000	59290	59316	-740685	26	2418

**Figure 2.21 Input of Beta-defensin in excel spreadsheet.**

+	305	16.4	0	1.8	NW_007281425.1:3080000-3880000	4499	4554	-795447	55	120
+	388	9.4	0	0	NW_007281425.1:3080000-3880000	4674	4726	-795275	52	544
+	1696	14.8	0.7	3.8	NW_007281425.1:3080000-3880000	5270	5570	-794431	300	3083
+	2453	3.8	0	0	NW_007281425.1:3080000-3880000	8653	8945	-791056	292	777
+	317	27.9	8.5	0	NW_007281425.1:3080000-3880000	9722	9850	-790151	128	2287

**Figure 2.22 Gap exceeding 3000bp in cluster region between whole Beta-defensin genes.**

By identifying gaps between the known genes of more than 3000bp (figure 2.22) and using BLAST against these sequences should give confidence that all the potential beta-defensins in the cluster had been identified and that a working method for looking into other species/DNA sequences can be employed.

## **2.6 Additional changes during development of methodology pipeline.**

During the study and development of this protocol several notable changes occurred to the methodology. In 2019, van Hoek *et al.* wrote a paper describing the cluster found within the Komodo dragon genome. The paper described that the beta-defensin cluster is flanked with Cathepsin B (CTSB) and that the other flanking gene is either Translocation associated membrane protein 2 (TRAM2) or Exportin 1 (XPO1). The characterised CTSB and TRAM2/XPO1 genes were downloaded from the NCBI website and were used for the initial searches to map out the potential region for further probing with the concatemer approach as outlined in section 2.2. Furthermore GIRI, as of May 20<sup>th</sup> 2019, rescinded their working agreement with RepeatMasker to utilise the RepBase database, which had an impact in the repeat masking ability of their website as this only allowed the use of searching DNA sequences with the Dfam database which was not as comprehensive as RepBase. However, with a local version of RepeatMasker downloads and the acquisition of the RepBase database for this program the analysis was able to be performed without the need for the online RepeatMasker database.

As reptilian genomes become more assembled into chromosomal builds these were then used for the final analyses within this manuscript.

One thing to note because of the variability of the beta-defensins when looking at the algorithm parameters, uncheck the box that masks for low complexity regions. This will allow the search programme to show results that may not be put forward otherwise. Such as the nature of beta-defensins when looking for searches the only thing that may give you a hit is the conserved cysteine motif, common to all beta-defensins and by not unchecking this you are limiting the number of matches you may receive from the search.

### ***3.0 Evaluation and comparison of search strategy from Santana, F. L. et al. (2021)***

The search methodology for the beta-defensins described in Santana *et al* (2021) was different to this work whereby the use of Markov models was employed. There are some differences between the sequences established by the methods from *Santana, F. L. et al. (2021)* and the sequences in this work. An evaluation was performed when the sequences were compared when two different search strategies were employed. The difference in the sequences from the *C. porosus* cluster region that were identified in the Santana paper have been described in tables 3.6 and 3.7. In their findings, four sequences show differences in the identified cluster region. CpoBD8 shows the amino acid sequence terminating with a valine residue. In the DNA sequence analysed in this work a stop codon was identified next to the codon for this residue and a potential third exon was identified in the genomic sequence used in this work. In cpoBD10 the end sequence shows that the last four residues are VPLG. This is different from what was found in this work at the genomic sequence shows that there is a stop codon before these residues in the sequence. CpoBD9 is said to overlap cpoBD2 in the Santana analysis, however, when analysing the sequence for cpoBD9 and using search tools of BLASTn against the genome and cluster region showed that this was not present in either DNA sequence, therefore it can be deduced that this does not overlap with cpoBD2. Finally, the splice site between first and second exon in cpoBD18 shows differences. When looking at the genomic sequence there is no donor splice site at the position which is given in their sequence.

The *A. mississippiensis* sequences in the Santana paper also showed a few differences with a total of three within the cluster region in comparison to defensins identified in this work. In amiBD5 there is a stop codon after the arginine residue at the end of the sequence and their sequences suggests that the sequence end with HHRTRD. AmiBD16 shows the presence of the first exon sequences, but the second exon is absent from the Genomic and cluster region. AmiBD3 and AmiBD18 shows different splice sites which are not present in the genomic sequence used by this analysis. These differences could be accounted from by their use of a previous version of the genome sequence.

	CROCODYLUS POROSUS		
Beta-defensin in Santana et al 2021	Sequence	Corresponding BD in this work	Notes on differences.
cpoBD1	MRILLLSALLFLVLQVQAQHKAEQAQDPALQDEAEAVMAAPENTPISRSSCRRSRGATCRVGFCEGELRLGSCAFLRPCKELPGL	CPBD9	
cpoBD2	MKLLFLLGVTTLVFQAQAQDVVVAQDKAEPQDLDEMEEEAEVMEAQDAAGMDFPGLNLGESPAPHCRWRRRGICRPTHCKKNDPNCRYNPCRFRQERIVGWCLSSHVCCVKAK	CPBD7	
cpoBD3	MKFLYLLFGVFLVLQPQAQDIQAQNKAEIQELNPAQPRRRKFCFRRGVCKSRCSRNEDSARRCRNRQHCCIKRRH		Not in Genome Assembly or contain between CTSB-TRAM2, Not included in the figure in Santana paper.
cpoBD4	MKIVYLLGVAFVLSQTEAQDVVVTQGEAEAQDLDEMDEEAKDNAMEAEYAARMGSPDVKPEFPVVCRIILGVCRFRCRKNERTIGSCSSRACCKRR		Not in between CTSB-TRAM2. Not included in the figure in Santana paper
cpoBD5	MRTLKLLFAVSLFMVQIAPGFFQIYGNTKLCKLNGGSCFLRSCPRKFVSFGTCTRECMCCIR	CPBD11	
cpoBD8	MRVLYLLFTVSILMLQLAAGFPKIGYFHCRSQNGNCYQYACPPNTKYIGSCNKLGNCCQRV	CPBD12	No stop codon next to codon for V in sequence before poly a signal. Potential third exon found in this analysis
cpoBD9	MKLLYLLGVATLVFQAQAQVTVVAQGEAEPQDLGEKQEAEDNIMEAEDAGYKGSADLKPLPSPLWCGWKGGYCRHHCKKEERTGWCTTNYVCCH		Not overlapping with cpobd2. Not in genome assembly, Not between CTSB-TRAM2
cpoBD10	MRFLYLLAVLFFLFQVSSGFVDVAPADTVACRNQGNFCRLGTCPTTFEGTGTCCNNGALLCCSKVPGL	CPBD10	Not VPLG at the end as stop codon is after K.
cpoBD12	MAGKRMLWFAAFLLAVPGNAQGSKHVCRTAGGQCRMGICLSGEVRIGDCFIPVILCCKYPVRKETGELQGGA	CPBD2	
cpoBD13	MRLLYLLFAAVMLLFLQAVPANGSYSTLQCRNHHGCRRLCFHGEQWIGNCNGRHQHCK	CPBD1	
cpoBD14	MRILYLLALLFLCQALADTLCTKNNGTCAFMLCPIFMKAIGTCYDGAACKCRRCI	CPBD13	
cpoBD15	MRILYLLFAVLLFVLQAAPGQPSRSLDRGGRCIRYNTCHPNLIINARCPHQTVCCRRR	CPBD6	
cpoBD17	MKILYLLVGLFLFLQAASGLGRCNLLNGVCRHTLCHSLEKYVGRCHRGLRNCCVDDYVLKYKM	CPBD14	
cpoBD18	MKLLYLLSVAFLVFQQAQDLKPHGSPDCHRKLKICRHVFCNLFEITIGYCNRRHHVCCRRWI	CPBD8	Spice site on this wrong as no GT in donor site in genome sequence corresponding to this sequence.
cpoBD19	TRIFLLAVLFFFHAAHPGHGQYHDKDRGGDCILHDTCLSSGEVIYAPCPRWLICRRRLR		No initiating Methionine
cpoBD20	MMKFFHLLALLFGIFLATTANGQRATRYVNHCLQKGGTCRYDDCEAGEEQIGTCYRQTMVCCRDEE	CPBD3	
cpoBD21	MKSLYLIALALFFSQVVPNGPLPILSFLQCLNLQGTCLLVGFCNGITIRLLGCDCTP	CPBD5	
cpoBD23	LRSLFLLFAVAFLLFQAAPPEEASPCRSFGDHCINWNERCRSGRFLAVPCPFRKRCCKS		No initiating Methionine

\*CPBD4 was determined in this work and not found by Santana et al 2021 methods.

**Table 2.3 Comparisons of *C. porosus* sequences showing differences identified in Santana et al. 2021.**

	ALLIGATOR MISSISSIPPIENSIS		
Beta-defensin in Santana et al 2021	Sequence	Corresponding BD in this work	Notes on differences
AmiBD1	MRVLLLLFALLFLVFQVQAQHQAEQAQDPALQDEAEAVMAAPENTPISRSNCKRSGATCRVGFCEGGEIKLGSCAFLRPCCKEPLGL	AMBD13	
AmiBD2	MKLLFLLLVGVTTLVFQAAQADVVVAQDEAEQDLGEMEEEEAEVMEAEADATGMDFPGPKLGESPAHCRWKRGVCRRTTHCKRNDNRNCRHTPCKPAERIIGWCLSTYVCCRKAY	AMBD8	
AmiBD3	MKFLYLLFGVAFVLVLTQAQDIQAQDKAEIQELNHPAQP RRRKFCRSRQGVCKPRCSGNENSSRRCRNHQRCCVKRRQ	AMBD12	Different splice site
AmiBD4	MKIYLLLVGAVFLVSQAQAQDVVAQDEAEQAQDLDDIDEAQAQDNAMAEYAATMGSPDVKPKQEPVVCVLLGVCRPFRCRLNERTIGSCSSNHACCKRY	AMBD11	
AmiBD5	MRTL YLLFAVSLFMVQIAPGFFQIYWNTKLCCLNGGSCFLRSCPRQFVFSFGTCTQECMCCIRHRTRD	AMBD15	Stop codon after R no sequences corresponding to HRTRD before poly a signal
AmiBD6	MKTPCLLFALVLLVLIHQAMPNPVGEKDPQKEADTWDEVEDDVGEEGDVEAQGR*GENSPMICGFSGGSCRTGCSSNEVMAGKCYGSLCCIPR		
AmiBD7	MKTPCLLFALVLLVHVQAMPNPVGEKQPHKEADTWGVEDDASKAKGNVEAEGAGGENNPMVCSYGGSCRQRCIGHEVMVGKCYGTFCVHM	AMBD14	
AmiBD8	MRVLYLLFAVSLMSQLAAGFPQIGYFHCQQNKGCQFQHICPPNTKYIGSCKQLGNCCQRV	AMBD16	
AmiBD9	MKLLYLLLVATLVFQAQAQVTVA*GEAEPQDLGEMQEQAEDNVMDAEDADDKGSADL*PLASPLWCGWKGGYCRHHCKEKERTGLCTVNYVCL		
AmiBD11	MKLLYLLVGVAFVLVFTQAQDGAVAQDEAEQAQDLDEMEEEAEDEFVEAEDAAGMGSPELARKDRPCRKGLFCRPKCGQKEHVIGTCKGLICRIL	AMBD10	
AmiBD12	MAEKRMWLVFVAILLAVPGNAQGSKNVCRSAGGQCQMGTCLSGEVRIGDCFTPVILCCKYLARKTPGELQGGA	AMBD2	
AmiBD13	MRILYLLFAAVMILFLQAVPAKGSYYSTLQCRNNHGHCRRLCFHRERWIGNCNGGHQHCK	AMBD1	
AmiBD14	MRILYLLLALLFLLCQALADTLTCTKNGTCSFMLCPIFMKAIGSCYDGAACKRRCI	AMBD17	
AmiBD15	MRILYLLFAVFLVLLQVAPGQSYRECRNRGGECRPHGSGHPGVSIVPVRCPHRTVCCR	AMBD7	
AmiBD16	MKTPCLLFALVLLVHVQAMPNPVGENDPQVEADTWDEVEDDAGEAEGDVEAEGAGGENSPMICGFSGGSCRTVCLISEVMAGKCYSSYLCLLPR		No second exon even with blast of DNA and AA sequences in scaffold and in the genome sequences. Could be due to more updated version of genome used for this analysis.
AmiBD17	MKILYLLVLGLFLFQAASGLGRNLLNGVCRHTLCHSLEKYIGRCHRGLRNCCVDDYVLKYKM	AMBD18	
AmiBD18	MKLLYFLSVAFLVFQAQAQALPKQGSPTDCHRQLGVCRSFLCFFFTTIGSCNRHQVCCRRI	AMBD9	Different splice site
AmiBD19	GIFQLLFHFIFLVAGHQEYHDCNRRGGDCILHDTCLSTGEIYAPCPRWLICCKRLR		No initiating methionine
AmiBD20	MMKFFYLLLVLFGLFATTANGQRASRYVNHCLQKGGTCRYDDCEAGEEQIGTCYRQTMVCCRDEE	AMBD3	
AmiBD21	MKSLYVILAVLFFSQVVPNGLPILSLIQCLNLGGICLISVSLCDGVTIRLLGCNCCSR	AMBD5	
AmiBD23	TRSLFLLFAVAFLLQAAPEEIVPSRFSGGYCIWNWERCRSGHFLVALCPFRKCKCS		No initiating methionine

**Table 2.4 Comparisons of *A. mississippiensis* sequences showing differences identified in Santana et al. 2021**



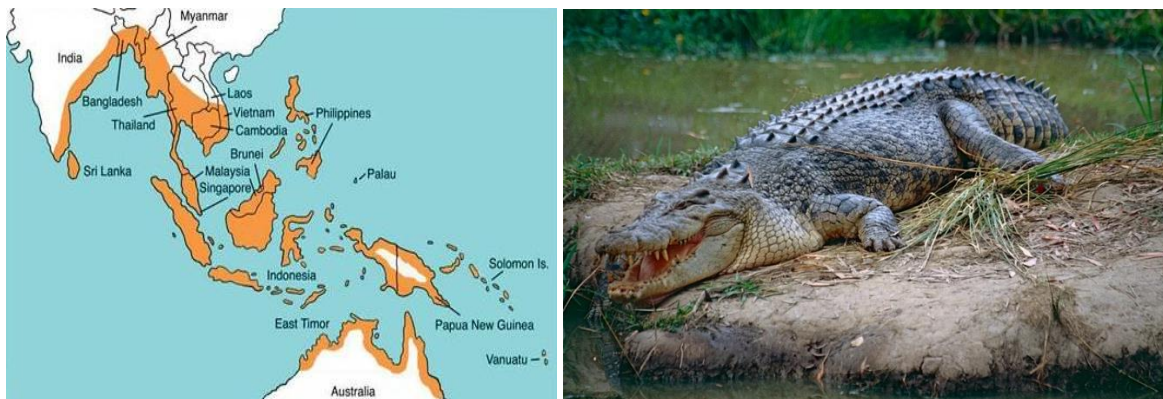
## Chapter 3 - Crocodylia

### 3. Aims

In this chapter two different orders of *Crocodylia* will be explored and the differences in number of beta-defensin genes present, physical properties of the peptide and genomic organisation will be discussed. The two species of crocodylia are the Saltwater crocodile, *Crocodylus porosus* and the American Alligator, *Alligator mississippiensis*.

**Note: Some of the sequences outlined in this chapter have previously been reported in Santana, F. L. et al. (2021) 'Reptilian  $\beta$ -defensins: Expanding the repertoire of known crocodylian peptides.'**

### 3.1 *Crocodylus porosus* - Salt Water Alligator



**Distribution map of *Crocodylus porosus* (Webb et. el. (2010) and image**

**(<https://www.flickr.com/photos/berniedup/10106331165/>)**

The Saltwater Crocodile is the largest living reptile and crocodylian known to science. It resides in the family Crocodylidae and genus *Crocodylus*. *Crocodylus porosus* was the scientific name proposed by Johann Gottlob Theaenus Schneider (Schneider 1801) and several other species were described over the years, however it is now considered a monotypic species. Their habitat is saltwater wetlands, and its distribution is from India's east coast across Southeast Asia to as far as Australia and South Pacific islands. Males can grow up to 6m in length and weigh up to 1000-1300kg with females being much smaller at almost

half their size. The saltwater crocodile is a large opportunistic apex predator. It is capable of hunting almost all animals that enter its territory including other apex predators such as sharks. It has been known to be responsible for several human deaths. The saltwater crocodile has a wide snout and has a pair of ridges that run from the eyes to the end of the snout. Its scales are oval and its scutes are often absent which is an unwelcome advantage when identifying illicit skins from illegal hunting. They have a large broad body which contrasts with other crocodiles which often causes mistakes in identity leading to them being wrongly identified as an alligator. They possess salt glands which allows them to survive and inhabit salt water and is feature alligators do not have. Saltwater crocodiles mate in the wet season and lay eggs in a nest consisting of mud and vegetation. The females guard the eggs from predators. The species is considered of minimal concern for extinction but is protected from the effects of international trade under the Convention on International Trade in Endangered Species (CITES). It is often hunted for their skin meat and eggs.

### **3.1.1 Data mining and cluster assembly**

The genomic sequencing data that was used for this analysis was obtained from The National Centre for Biotechnology Information (NCBI) genome assembly database. The genome was chosen as it was at scaffold level of genome assembly however the scaffold was large enough to allow the full cluster to be determined. The GenBank assembly accession number is GCA\_001723895.1 and was submitted to the database on 13/09/2016. The cluster region search was started with downloading the sequences for CTSB and TRAM2 and using tBLASTn program to search for the region in which the cluster may reside. The region was found to be on scaffold NW\_017728918.1 between locations 5774492-6109160, total length being approximately 334kb long. This region was then masked for repetitive sequences using RepeatMasker program to remove the repeat sequences from the DNA sequence. Once again this was translated into a 6-frame output so potential matches from the query search could be highlighted and further analysed.

The Beta-defensin sequences from the Green Lizard (*Anolis carolinensis*) were obtained from Dalla Valle *et al.* (2012). In supplementary Table 1 Dalla Valle *et.al* give a list of accession numbers that can be accessed via the NCBI website. These were put into a concatenation for

the query sequences in which the BLAST searching would be performed. Along with the sequences obtained from the Dalla Valle paper, sequences already identified from *Alligator mississippiensis* from the protein database at <https://pfam.xfam.org> were also put into a search query concatenation.

The tBLASTn program was applied against this region and highlighted on 6-frame output. This process, however, does not acquire all the exons and therefore other approaches were employed. Gene finding programs GENSCAN (Burge and Karlin 1997) and FGENESH (Solovyev *et al.* 2006) were employed to search for putative exons that were not initially found with the BLAST approach. Finally, regions of more than 3000bp of the repeats determined in the repeatmasker analysis but not in the vicinity of already resolved exons can then be searched to exclude all potential regions where Beta-defensin exons may reside.

Splice site prediction was finalised using the online server by the Berkeley Drosophila Genome Project (Reese *et al.* 1997) and amino translations were ascertained from the DNA sequences of potential exons. Finally, iterative searches were performed using the newly identified beta-defensins against the cluster region.

### **3.1.2 Cluster organisation and Beta-defensin sequences**

A total of 14 beta-defensins were identified (Appendix 4.4) within this region and were numbered according to the position on the chromosome starting from the nearest gene to CTSB, in this case CrPBD. This naming method was different to that of the Santana paper as their naming nomenclature follow that of orthologues found in birds. Relative positions and genomic organisation along the DNA region are depicted in figure 3.1. Positions of each exon and sizes are available in appendix 4.1.



**Figure 3.1 Genomic organisation of the *C. porosus* Beta-defensin cluster.**

Each vertical line represents 50kb along the chromosome. The blue blocks represent the full genes, and the arrows dictate their orientation. The size of the squares (genes) or the distance between them are only representative and are not proportional to their actual size in the genome.

All the beta-defensins identified show the classical structure and consist of two exons except for CrPBD12. Exon 1 encodes a conserved signal peptide followed by the second exon encoding the mature peptide. The conserved defensin motif is present with common 6 cysteine domain and a glycine located in the N-terminal region, two positions upstream from the second cysteine, while another glycine is placed two or three positions upstream from the fourth cysteine with the rest of the amino acids being less conserved but show similarities where the genes have recently duplicated. This is observed in the multiple sequence alignment showing conservation motif (figure 3.2). Three of the beta-defensins identified in this work have a long anionic pro-domain, and this has been described (Michaelson *et al* 1992) as a mechanism in which the pro-domain counterbalances the cationic charge of the active Beta-defensin during synthesis. Table 3.1 shows the charges between the long pro-domain beta-defensins minus the signal sequences and then the charge of the 2<sup>nd</sup> exon, which may closely represent the mature active form.

### **3.1.3 Physical Properties**

Each beta-defensin gene identified in this genome possesses a conserved signal peptide, and this was confirmed using SignalIP – 5.0 server (Almagro Armenteros *et al* 2019) (appendix 4.3). There is a wide range of charges and some of the beta-defensins in this cluster are anionic although most of the beta-defensins are cationic (Appendix 4.2). One such defensin, CrPBD7 has a charge of -5 but this is similar to what was found by Tang *et al* (2018) in that it has a long anionic pro-domain and this may serve to balance the charge of the defensin before

undergoing further post translational modifications to produce the active mature peptide. Table 3.1 shows the charges between the long pro-domain beta-defensins minus the signal sequences and then the charge of the second exon, which may closely represent the mature active form.

GENE	Long pro-domain mature peptide			Second Exon		
	pI	Net Charge	Mr	pI	Net Charge	Mr
CPBD7	5.11	-5	10792	9.45	8	6428
CPBD8	5.64	-4	9318	9.18	5	5121
CPBD9	5.74	-1	7396	9.25	5	4707

**Table 3.1 Charge differences between longer pro-domain peptides and the second exon for *C. porosus*.**

*Isoelectric point and molecular mass included.*

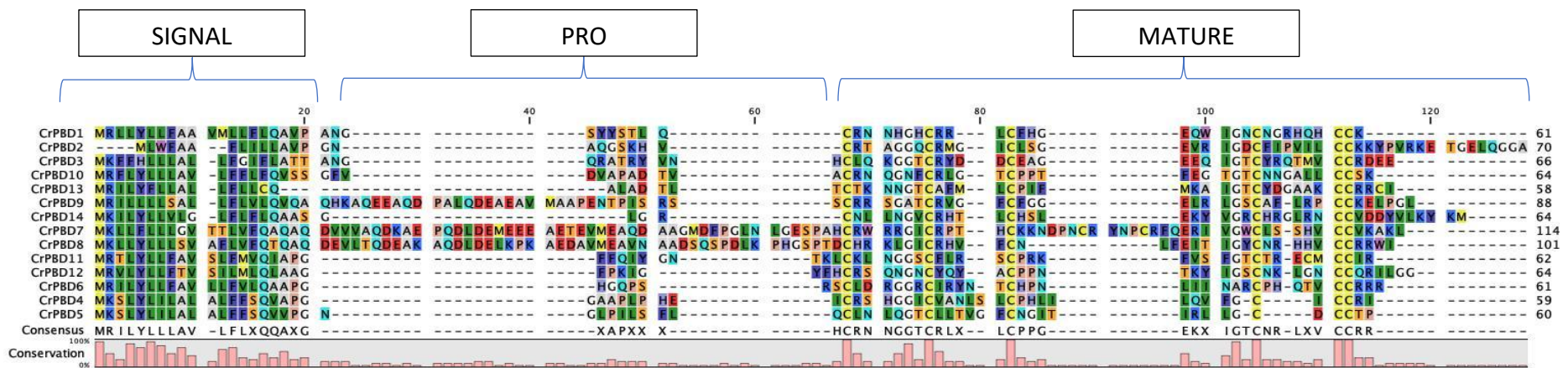
### 3.1.4 Selection Analyses

Multiple sequence alignments were produced in CLUSTALX (Larkin *et al* 2007) and Codon alignments subsequently produced using the PAL2NAL server (Suyama *et al* 2006). These codon alignments were the used in pairwise comparisons between nucleotide sequences, the number of synonymous substitutions per synonymous site ( $dS$ ) and the number of nonsynonymous substitutions per nonsynonymous site ( $dN$ ) were estimated using the SNAP v.2.1.1 program at <http://www.hiv.lanl.gov> which implements Nei and Gojobori (1986) method (Korber 2000). The proportion of observed synonymous and nonsynonymous substitutions were plotted against each other (figures 3.4 A & B). Viewing the distributions between the signal peptide and the second exon there are slight differences on the distribution of the points. The signal peptide shows a greater degree of points distributed towards synonymous substitutions showing a level of conservation between codons across the gene implying that it is undergoing possible purifying selection pressures. However, the second exon shows that the distribution is closer to  $dS=dN$  but still showing a slight purifying selection. This is most likely due to the number of paralogues having homology within the cluster. When observing the second exons within the whole cluster you may expect there to be a greater degree of nonsynonymous substitutions due the variation of amino acid

sequences present, therefore a site-wise analysis was performed to gain a better picture of the evolutionary dynamics within the individual sites within the gene.

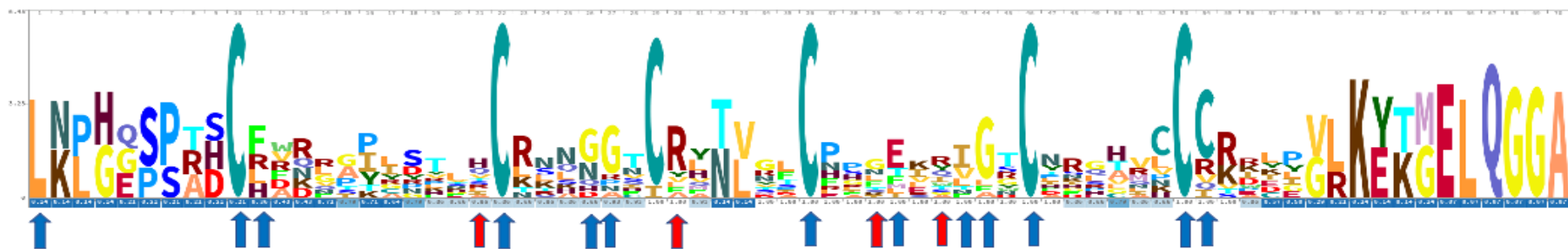
### **3.1.5 Repeat Sequence landscape**

Repeat masker was performed using query species database set to tetrapod. The *C. porosus* defensin cluster region had 37.23% bases masked with the predominant repeat elements being retroelements at 60.5% of bases masked. LINES were around 78% of the retroelements and CR1 LINE being the most abundant at 88.6% of the LINES present. LTR elements accounted for 37.3% of the retroelements. Around 37% of the repeat sequences were DNA transposons with hobo-Activator and Tourist/harbinger being the most abundant (table3.2).



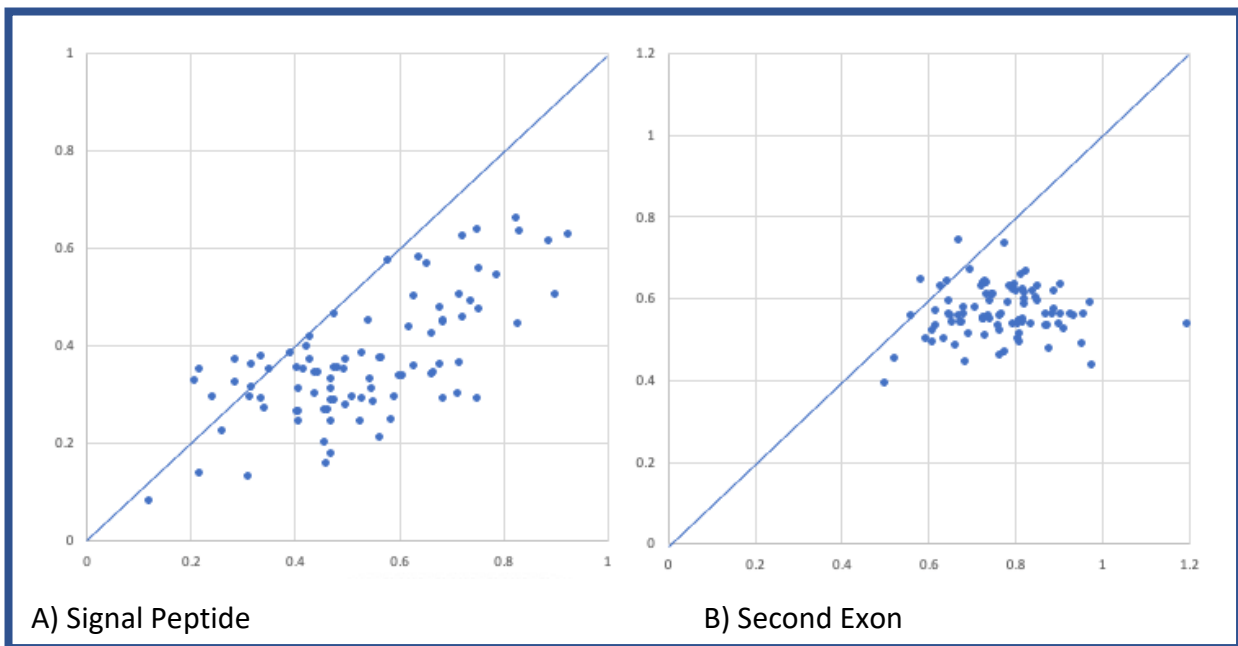
**Figure 3.2 Multiple sequence alignment of *C. porosus* beta-defensins cluster.**

Produced using Clustal X. Conservation of amino acids is shown in the legend underneath and show the typical conserved signal peptide at the start of the gene and in the mature peptide showing the 6 conserved cysteine residues along with glycine residues. Signal, Pro-peptide and Mature regions are also shown by the parentheses.



**Figure 3.3 Amino acid sequence logo of second exon.**

Sites which are undergoing positive selection (red arrow) by one of more tests and purifying selection (blue arrow) tested by FEL, FUBAR and MEME in HYPHY. Logo produced on Skylign.org



**Figure 3.4. Ratio of synonymous and nonsynonymous substitutions.**

Within the signal peptide (A) and the second exon peptide (B). Graphs show synonymous ( $d_N$ ) on the x axis and nonsynonymous ( $d_S$ ) on the y axis. The diagonal lines represent  $d_N = d_S$  and is given for estimating selection pressures; dots above and below this line represent positive and purifying selection, respectively.

Within the second exon there are four residues within the amino acid sequence that are undergoing positive selection and 13 residues undergoing negative/purifying selection. The positions that are undergoing positive selection are located between the positions that are undergoing negative selection. The negatively selected amino acids are shown to be residues that are common to beta-defensins. These are the 6 cysteines that make up the covalent bonding that is seen throughout the defensin class along with the glycine residues notably the GxC residues and the second and fourth cysteine residues that make up the beta sheets integral to its structure (Tu *et al* 2015). However, the residues that are undergoing positive selection are located in the regions that contribute to the bends around these beta sheets and sited on the outside of the peptide.



```

=====
file name: Crocodylus porosus_REV_COMP.fa
sequences: 1
total length: 334669 bp (326381 bp excl N/X-runs)
GC level: 49.29 %
bases masked: 124583 bp ( 37.23 %)
=====

```

	number of elements*	length occupied	percentage of sequence
Retroelements	235	75484 bp	22.55 %
SINEs:	26	3653 bp	1.09 %
Penelope	19	3268 bp	0.98 %
LINEs:	173	58946 bp	17.61 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	132	50998 bp	15.24 %
R1/LOA/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	0	0 bp	0.00 %
RTE/Bov-B	8	1882 bp	0.56 %
L1/CIN4	4	850 bp	0.25 %
LTR elements:	36	12885 bp	3.85 %
BEL/Pao	0	0 bp	0.00 %
Ty1/Copia	0	0 bp	0.00 %
Gypsy/DIRS1	14	8014 bp	2.39 %
Retroviral	15	2892 bp	0.86 %
DNA transposons	212	46517 bp	13.90 %
hobo-Activator	124	27531 bp	8.23 %
Tc1-IS630-Pogo	3	943 bp	0.28 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	71	16725 bp	5.00 %
Other (Mirage, P-element, Transib)	0	0 bp	0.00 %
Rolling-circles	3	77 bp	0.02 %
Unclassified:	5	848 bp	0.25 %
Total interspersed repeats:		122849 bp	36.71 %
Small RNA:	5	547 bp	0.16 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	31	1327 bp	0.40 %
Low complexity:	5	174 bp	0.05 %

```

=====
* most repeats fragmented by insertions or deletions
  have been counted as one element

The query species was assumed to be tetrapods
RepeatMasker version 4.1.2-p1 , default mode

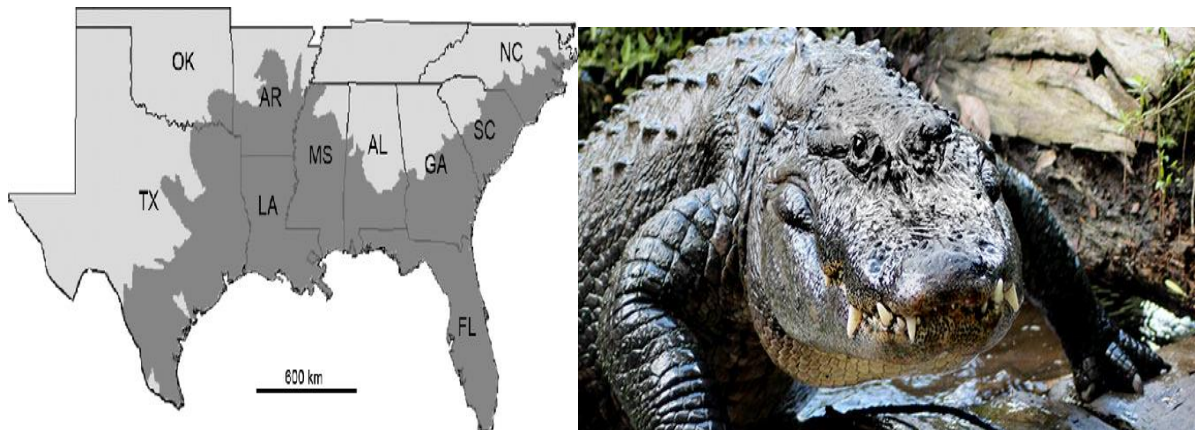
run with rmblastn version 2.2.27+
FamDB: CONS-Dfam_withRBRM_3.3
=====

```

**Table 3.2 Repeat masker summary for *C. porosus***

Displaying the different repeat sequences within the *C. porosus* cluster region. The tetrapod database was used as a reference for the repeat sequence matches in RepeatMasker program.

### 3.2 The American Alligator - *Alligator Mississippiensis*



Distribution map (Ferraro and Binetti (2014)). Photo <https://www.nwf.org/Educational-Resources/Wildlife-Guide/Reptiles/American-Alligator>

The American Alligator, *Alligator mississippiensis*, is one of two extant alligator species and is possibly the largest species in the family Alligatoridae as it is larger than the only other species of alligator, the Chinese alligator: *Alligator sinensis*. Its alternative name is the Common Alligator and is native to the South-eastern United States. Their distribution is South Carolina, Georgia, Florida, Louisiana, Alabama, Mississippi, Arkansas, and Texas. Louisiana has the largest population of all the U.S. states and sometimes they can be found in Mexico. American Alligators inhabit slow-moving rivers, swamps marshes and lakes. They prefer fresh water rather than salt water as compared to their American cousins, the American Crocodile, do not have functioning salt glands on their tongue. The American Alligator is an apex predator and consumes fish, amphibians, reptiles, mammals, and birds, given the chance. Adult males can reach 3.4-4.8m in length and can weight up to 450kg but there have been unverified sightings of alligators of up to 5.5m in size and weighing a 1000kg. It is identified from the American crocodile by its broader snout with overlapping jaws and darker skin colour. It is also tolerant to cold temperatures with it being able to survive at temperatures as low as 7°C and can go into brumation when the water temperatures go below freezing. Their breeding season begins in spring and after mating the females can lay up to 20-50 white eggs in which she covers with vegetation within their nests. The incubation period for the eggs is 65 days and eggs that hatch above 34°C will hatch as males and females will hatch when the temperature is below 30°C. The American Alligator is currently of at least concern

by the IUCN Red list even though they were extensively hunted from the 1800s to the mid-1900s. American Alligators trade is regulated under the Convention of International Trade in Endangered Species (CITES)

### **3.2.1 Data mining and cluster assembly**

As with *C. porosus*, the genomic sequencing data utilised for this analysis was obtained from The National Centre for Biotechnology Information (NCBI) genome assembly database. The genome was chosen as it was at scaffold level of genome assembly however the cluster region was separated on two different scaffolds but by joining these the full cluster region was able to be determined. The GenBank assembly accession number is GCA\_000281125.4 and was submitted to the database on 28/03/2016. CTSB and TRAM2 amino acid sequences were downloaded from the orthologue list on the NCBI database and were used to search the genome assembly to establish the cluster region. CTSB resided on scaffold NW\_017709158.1 starting at position 259102. The reverse complement of position 1-259102 was used for the start of the cluster. TRAM 2 resided on scaffold NW\_017710918.1 at position 2229362. This position through to the end of the scaffold was taken and the reverse complement used to produce the rest of the cluster region. The two scaffolds were joined to produce the cluster region DNA sequences for further analysis. The total length of the construct was approximately 395kb. This region was masked using the RepeatMasker.org server (Smit *et al.* 2006) to remove the repeat sequences from the DNA sequence. The was then translated into a 6-frame output using EMBOSS Sixpack program on The European Bioinformatics Institute (EMBL-EBI) website and this was utilised to highlight potential matches.

Unlike the approach of using concatemers outlined above, a FASTA file of the DNA coding sequences from *C. porosus* was employed to run the search of the potential beta-defensins that reside within the region that was uncovered between CTSB and TRAM2. The DNA coding sequences from *C. porosus* were used as a query against the genome using the BLASTn program and due to the orthology of these sequences as the initial *A. mississippiensis* sequences were found to be highly homologous. This DNA coding search approach identified most of the genes present but to have confidence that all the genes were uncovered the concatemer approach was applied using the *C. porosus* amino acid sequences followed by the gene finding programs GENSCAN (Burge and Karlin 1997) and FGENESH (Solovyev *et al.* 2006)

and finally searching the regions of more than 3000bp between repeat sequences shown by the running of the RepeatMasker program.

Splice site prediction was finalised using the online server by the Berkeley Drosophila Genome Project (Reese *et al.* 1997) and amino translations were ascertained from the DNA sequences of potential exons. Finally, iterative searches were performed using the newly identified beta-defensins against the cluster region.

### 3.2.2 Cluster organisation and Beta-defensin sequences

A total of 18 beta-defensins were identified (Appendix 4.1) within this region and were numbered according to the position on the chromosome starting from the nearest gene to CTSB, in this case AMBD. AMBD1-13 were identified on scaffold NW\_017709158.1 and AMBD14-18 resided on NW\_017710918.1. Relative positions and genomic organisation along the DNA region are depicted in figure 3.5. Positions of each exon and sizes are available in appendix 4.4.



**Figure 3.5 Genomic organisation of the *A. mississippiensis* Beta-defensin cluster.**

*Each vertical line represents 50kb along the chromosome. The blue blocks represent the full genes, and the arrows dictate their orientation. The size of the squares (genes) or the distance between them are only representative and are not proportional to their actual size in the genome. Double diagonal line showing the end of one scaffold and start of the next.*

All the beta-defensins identified show the typical structure and consisting of two exons. Exon 1 encodes a conserved signal peptide followed by the second exon encoding the mature antimicrobial peptide. The conserved defensin motif is present with common 6 cysteine domain and a glycine located in the N-terminal region, two positions upstream from the second cysteine, while another glycine is placed two or three positions upstream from the fourth cysteine with the rest of the amino acids being less conserved but show similarities

where the genes have recently duplicated. This is observed in the multiple sequence alignment showing conservation motif (figure 3.7). Three of the beta-defensins identified in this work have a long anionic pro-domain, and this has been described (Michaelson *et al* 1992) as a mechanism in which the pro-domain counterbalances the cationic charge of the active Beta-defensin during synthesis.

### 3.2.3 Physical Properties

Each beta-defensin gene identified in this genome possesses a conserved signal peptide, and this was confirmed using Signal IP – 5.0 server (Almagro Armenteros *et al* 2019) (Appendix 4.6). There is a wide range of charges and some of the beta-defensins in this cluster are anionic although most of the beta-defensins are cationic (Appendix 4.5). One such defensin, CMBD20 has a charge of -7 and this concurs with findings in Crocodylia (Tang *et al* 2018) in that it has a long anionic pro-domain and this may serve as a way to balance the charge of the defensin before undergoing further post translational modifications to produce the active mature peptide. Table 3.4 shows the charges between the long pro-domain beta-defensins minus the signal sequences and then the charge of the 2<sup>nd</sup> exon, which may closely represent the mature active form.

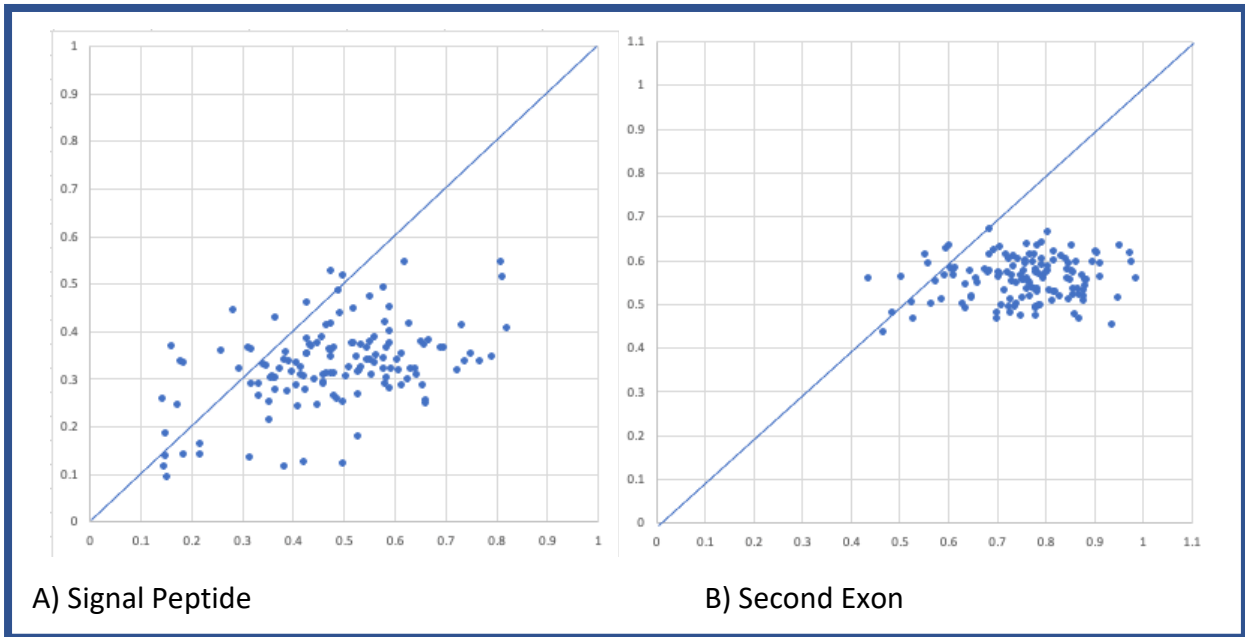
	Long pro-domain mature peptide			Second Exon		
	pI	Net Charge	Mr	pI	Net Charge	Mr
AMBD8	5.61	-3	10292	9.86	11	6672
AMBD9	4.97	-4	8957	8.98	4	5102
AMBD10	4.38	-8	9025	9.77	8	4483
AMBD11	4.55	-8	8594	9.15	5	4957
AMBD12	9.58	8	9399	11.38	12	5145
AMBD13	5.74	-1	7367	8.94	4	4807
AMBD14	5.17	-3	8105	7.79	1	4443

**Table 3.4 Charge differences between the longer pro-domain/mature peptides and the second exon for *A. mississippiensis*.**

*Isoelectric point and molecular mass included.*

### 3.2.4 Selection Analyses

Multiple sequence alignments were produced in CLUSTALX (Larkin *et al* 2007) and Codon alignments subsequently produced using the PAL2NAL server (Suyama *et al* 2006). These codon alignments were used in pairwise comparisons between nucleotide sequences, the number of synonymous substitutions per synonymous site ( $dS$ ) and the number of nonsynonymous substitutions per nonsynonymous site ( $dN$ ) were estimated as described in section 3.1.4. The proportion of observed synonymous and nonsynonymous substitutions were plotted against each other (figure 3.6 A and B). Viewing the distributions between the signal peptide and the second exon there are slight differences on the distribution of the points. The signal peptide shows a greater degree of points distributed towards synonymous substitutions showing a high level of conservation between codons across the gene implying that it is undergoing possible purifying selection pressures. However, the second exon shows that the distribution is closer to  $dS=dN$  but still showing a slight purifying selection. This is most likely down to the number of paralogues within the cluster having homology within the cluster. When observing the second exons within the whole cluster one may expect there to be a greater degree of nonsynonymous substitutions due to the variation of amino acid sequences present, therefore a site-wise analysis was performed to gain a better picture of the evolutionary dynamics within the individual sites within the gene.



**Figure 3.6 Ratio of synonymous and nonsynonymous substitutions in *A. mississippiensis*.**

Within the signal peptide (A) and the second exon peptide (B). Graphs show synonymous ( $d_N$ ) on the x axis and nonsynonymous ( $d_S$ ) on the y axis. The diagonal lines represent  $d_N = d_S$  and is given for estimating selection pressures; dots above and below this line represent positive and purifying selection, respectively.

Within the second exon there are 2 positions that are undergoing positive selection and 12 residues undergoing negative/purifying selection. The positive-selection positions are located between the negative-selection positions. The negatively pressured amino acids are shown to be residues that are common to beta-defensins. These are the 6 cysteines that make up the covalent bonding that is seen throughout the defensin class along with the glycine residues notably the GxC residues and the second and fourth cysteine residues that make up the beta sheets integral to its structure (Tu *et al* 2015). However, the residues that are undergoing positive selection are located in the regions that contribute to the bends around these beta sheets and sited on the outside.

```

=====
file name: Alligator_Mississippiensis_Contigs_full.fa
sequences: 1
total length: 395165 bp (385512 bp excl N/X-runs)
GC level: 48.34 %
bases masked: 157516 bp ( 39.86 %)
=====

```

	number of elements*	length occupied	percentage of sequence
Retroelements	274	100123 bp	25.34 %
SINEs:	19	2544 bp	0.64 %
Penelope	22	3302 bp	0.84 %
LINEs:	193	67485 bp	17.08 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	138	56566 bp	14.31 %
R1/L0A/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	1	19 bp	0.00 %
RTE/Bov-B	11	4078 bp	1.03 %
L1/CIN4	7	1329 bp	0.34 %
LTR elements:	62	30094 bp	7.62 %
BEL/Pao	0	0 bp	0.00 %
Ty1/Copia	0	0 bp	0.00 %
Gypsy/DIRS1	25	15653 bp	3.96 %
Retroviral	30	12122 bp	3.07 %
DNA transposons	246	53218 bp	13.47 %
hobo-Activator	140	31276 bp	7.91 %
Tc1-IS630-Pogo	5	1287 bp	0.33 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	79	18206 bp	4.61 %
Other (Mirage, P-element, Transib)	0	0 bp	0.00 %
Rolling-circles	2	62 bp	0.02 %
Unclassified:	4	1002 bp	0.25 %
Total interspersed repeats:		154343 bp	39.06 %
Small RNA:	3	284 bp	0.07 %
Satellites:	3	560 bp	0.14 %
Simple repeats:	59	2219 bp	0.56 %
Low complexity:	5	166 bp	0.04 %

```

=====
* most repeats fragmented by insertions or deletions
  have been counted as one element

The query species was assumed to be tetrapods
RepeatMasker version 4.1.2-p1 , default mode

run with rmblastn version 2.2.27+
FamDB: CONS-Dfam_withRBRM 3.3
=====

```

**Table 3.5 Repeat masker summary for *A. mississippiensis*.**

Displaying the different repeat sequences within the cluster region. The tetrapod database was used as a reference for the repeat sequence matches in RepeatMasker program.



### **3.2.5 Repeat Sequence landscape**

Repeat masker was performed using query species database set to tetrapod. The *A. mississippiensis* defensin cluster region had over all 36.86% bases masked with the predominant repeat elements being retroelements at 64.7% of bases masked. LINES were around 67.4% of the retroelements and CR1 LINE being the most abundant at 83.8% of the LINES present. LTR elements accounted for 30% of the retroelements. Around 33.7% of the repeat sequences were DNA transposons with hobo-Activator and Tourist/harbinger being the most abundant (table 3.5).

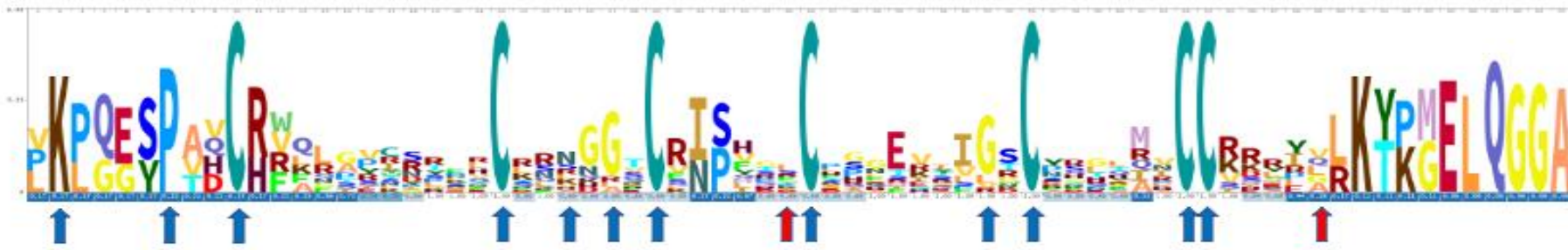
### **3.3 Conservation of synteny**

Synteny between these two species shows a high degree of homology not just in the cluster region DNA sequences but also in the distribution of the Beta-defensin genes along with their sequences. The cluster regions are both flanked by CTSB at the start of the region and TRAM2 on the other end. The dot plot of each DNA cluster region shows that sequences are highly similar with just two noticeable gaps which are accounted for by regions where gene had duplicated in the cluster region of *A. mississippiensis* (figure 3.9).



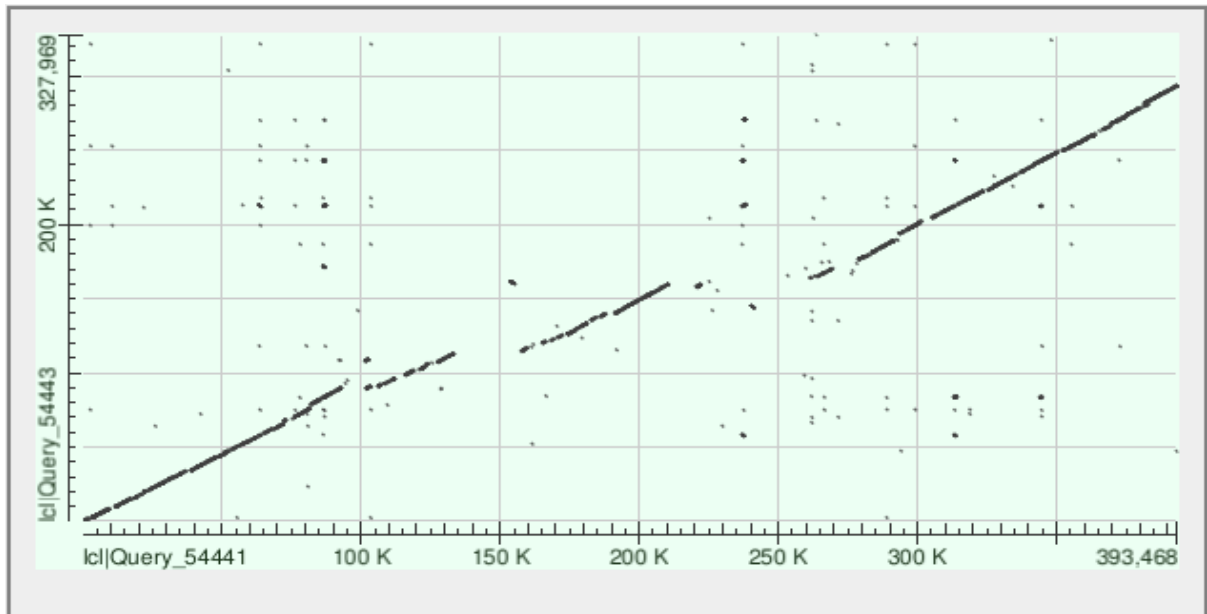
**Figure 3.7 Multiple sequence alignment of *A. mississippiensis* beta-defensin sequences.**

Produced using Clustal X. Conservation of amino acids is shown in the legend underneath and show the typical conserved signal peptide at the start of the gene and in the mature peptide showing the 6 conserved cysteine residues along with glycine residues. Signal, Pro-peptide and Mature regions are also shown by the parentheses.

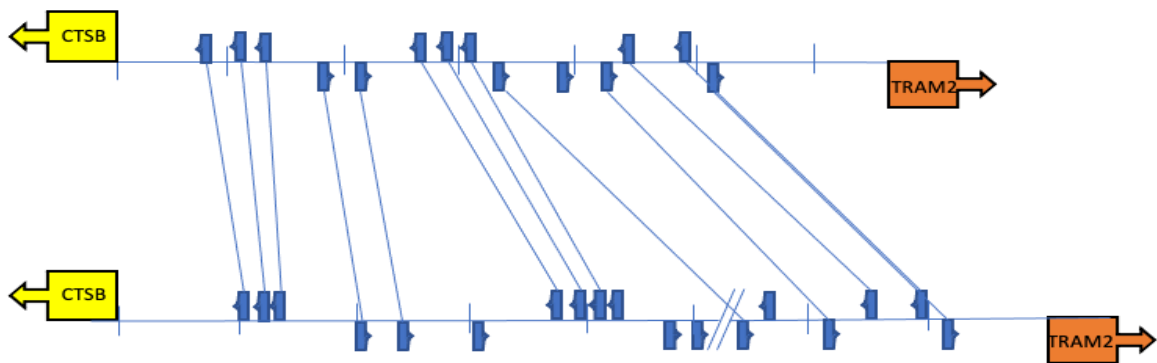


**Figure 3.8 Amino acid sequence Logo of the second exon in *A. mississippiensis*.**

Sites which are undergoing positive selection (red arrow) by one of more tests and purifying selection (blue arrow) tested by FEL, MEME and FUBAR in HYPHY. Logo produced on Skylign.org



A



B)

**Figure 3.9 Conservation of Synteny.**

- A) Dot plot comparison of both *A. mississippiensis* (x-axis) and *C. porosus* (y-axis). Areas of high homology can be observed by the diagonal line running upwards from left to right.
- B) Genomic organisation of both *C. porosus* (top) and *A. mississippiensis* (bottom). Lines represent orthologues of genes.

### **3.4 Summary**

The Saltwater Crocodile and the American alligator complete Beta-defensin clusters were identified in the genome assemblies and were found to reside between CTSB and TRAM2. There is a high degree of synteny between the two species in homology between gene sequences and genomic arrangement. The Saltwater Crocodile had a total of 14 complete genes identified and the American Alligator had 18 complete genes. Genes within this cluster shared similarities in structure and both possessed genes that have a long propiece in the primary amino acid sequences. This has been hypothesised to act as a charge balancer within the peptide. The genes within the cluster show a conserved signal peptide along with a more variable mature active peptide. The mature peptide shows several sites that are undergoing negative/purifying selection and few sites undergoing positive selection. The repeat sequence landscape shows a large percentage of the cluster region contains repetitive sequences.

## Chapter 4 - LIZARDS

### **4. Aims**

This chapter will focus on three lizard species that are native to Europe but also extend into the Eurasian continent. The numbers of genes present, genomic organisation and physical properties will be explored along with conservation of synteny analysis describing similarities and differences within the cluster.

The three species are *Podarcis muralis*, *Lacerta agilis* and *Zootoca vivipara*. The sequences described in this section were obtained by *in silico* means by the methods outlined in chapter 2.

### **4.1 *Podarcis muralis* – Common Wall Lizard.**



#### ***Distribution map and image of Podarcis Muralis.***

Photos obtained from - <https://www.eurolizards.com/lizards/podarcis-muralis/>

The Common Wall Lizard is a species of lizard with a large distribution in Europe as well as introduced populations in North America. There are several subspecies with the nominate subspecies occurring in the Balkan Peninsula. Their colourings are often brown with a dark brown stripe along the flank which occurs from the ear. Their underside is white or reddish. They can grow up to 20cm with the tail making up to two thirds of their total length. Their habitat prefers rocky environments including urban settings for which it can take shelter in the nooks and crannies of rocks and walls where they can conceal themselves easily. Their

diet is mainly insectivorous and when larger can handle smaller invertebrates. The female usually lays 2-6 eggs in one or more burrows.

#### **4.1.1 Data Mining and Cluster assembly**

The genomic sequence data was obtained from the NCBI genome assembly database. This species of lizard was chosen as the assembly was at chromosomal level and likely to have the fully assembled Beta-defensin cluster region that could be annotated and investigated further. The genome GenBank assembly number is GCA\_004329235.1 and was submitted to the database on 07/03/2019. These genes were then searched for in the genome to see if a region was identified. Using CTSB (cathepsin B) as the reference for the start of the analysis, the chicken CTSB amino acid sequence was downloaded from the orthologue list in an NCBI database gene search. Van Hoek (2019) states that either XPO1 or TRAM2 could be flanking the other end of the cluster so these were then used as a reference to establish the other end of this region using the tBLASTn program. It was found that the Beta-defensin cluster region resided between CTSB at the start of the cluster and XPO1 (exportin 1).

Using the amino acid sequences obtained whilst developing the search method and from the sources outlined therein, concatemers were produced for this analysis along with the partial sequences produced from the scaffold of the beta-defensins discovered in the *Chrysemys picta bellii*. These concatemers were then used as a query search within this region using the tBLASTn program on the BLAST server on the NCBI website. This region was masked for repeat sequences using RepeatMasker software using the tetrapod sequences within the database and translated into a 6-frame output using the EMBOSS sixpack program (EMBL-EBI) and used to highlight potential matches from the initial query searches. By using the sequences obtained in the method development and using a concatemer this produced a high 'hit rate' for potential matches but due to the size of the first exons in the Beta-defensin sequences the gene finding programs GENSCAN and FGENESH were employed. Finally, regions of more than 3000bp of the repeats determined in the repeat masker analysis but not in the vicinity of already resolved exons and downstream from the poly adenylation signal, can then be queried to exclude all potential regions where Beta-defensin exons may reside.

Splice sites were then predicted, and DNA coding sequences were translated into protein sequences.

#### 4.1.2 Cluster organisation and Beta-defensin sequences

A total of 80 genes were identified (appendix 1.1) spanning a region of approximately 2.05 Mb long (figure 4.1). The cluster was found on chromosome 3 - GenBank sequence CM014745.1 between locations 115993322-118049604 for which the reverse complement was used as to start the cluster at the CTSB gene. The naming of the genes used a prefix to the order number and was an abbreviation of the species name, in this case PMBD.



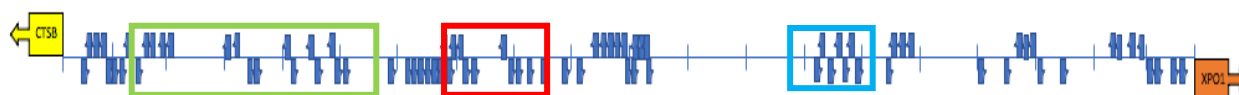
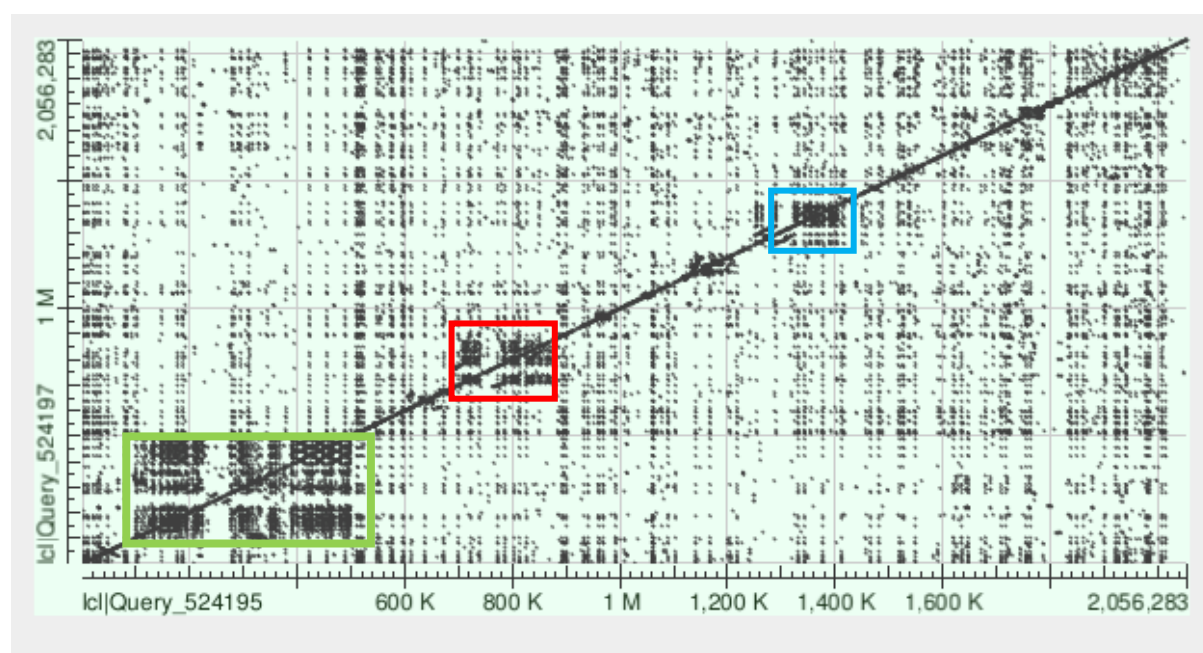
**Figure 4.1 Genomic organisation of the *P. muralis* Beta-defensin cluster.**

Each vertical line represents 100kbp along the chromosome. The blue blocks represent the full genes, and the arrows dictate their orientation. The size of the squares (genes) or the distance between them are only representative and are not proportional to their actual size in the genome.

The beta-defensins that were discovered in this analysis show the typical structure in that the gene consists of two exons. The first exon, a signal peptide followed by the second exon, the mature peptide that contains the defensin motif with common 6 cysteine domain and a glycine located in the N-terminal region, two positions upstream from the second cysteine, while another glycine is placed two or three positions upstream from the fourth cysteine. One interesting beta-defensin, BD72, shows a duplication of the mature peptide sequence (highlighted in pink in appendix) and has two of the common cysteine patterns present in the second exon. A similar peptide has been identified in birds, AvBD11, which has a double motif produced by the fusion of two exons (Guyot *et al.* 2020). Their study showed the peptide has multiple roles, such as a broad antimicrobial, antiparasitic activity and likely to have a part in how the embryo develops in the egg. Therefore, this peptide would be worth investigating further.



There are regions that contain high levels of duplicated genes, and these are reflected in the amino acid sequences shown highlighted in green, red and blue in the dot plot (figure 4.2) and highlighted corresponding coloured parentheses on the multiple sequence alignment (figure 4.3).



**Figure 4.2** Dot plot of the cluster region and genomic organisation of *P. muralis* genes.

Regions of high duplication highlighted with green, red and blue boxes. The dot plot was produced by using the genomic sequence against itself.

### 4.1.3 Physical Properties

Each beta-defensin gene identified in this genome possesses a conserved signal peptide, and this is seen in the sequence analyses using SignalIP – 5.0 server (Almagro Armenteros *et al* 2019) (appendix 1.3). There is a wide range of charges and some of the beta-defensins in this cluster are highly anionic which is contrary to the notion that beta-

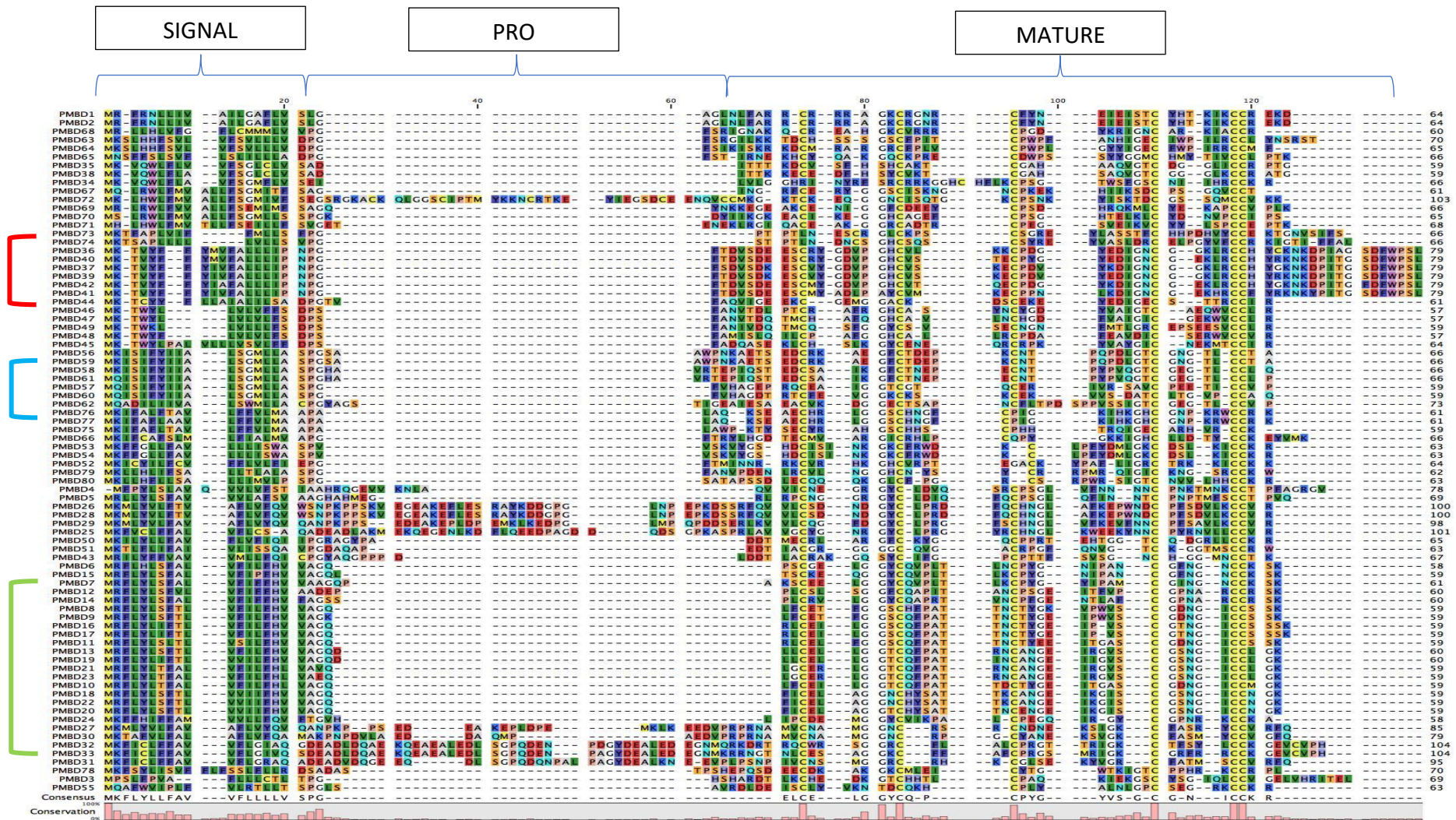


defensins are usually cationic. One such defensin, PMBD32 (Appendix 1.2), has a charge of -8 but this is similar to what was found in crocodylians (Tang *et al* 2018). It has a long anionic pro-domain, and this has been described (Michaelson *et al* 1992) as a mechanism in which the pro-domain counterbalances the cationic charge of the active Beta-defensin during synthesis. Table 4.1 Shows the charges between the long pro-domain beta-defensins minus the signal sequences and then the charge of the 2<sup>nd</sup> exon, which may closely represent the mature active form. From this, there is a difference in charge supporting the observations made by Michaelson *et al* (1992). Tang *et al* (2018) also states that this expression pattern has been observed in the small intestine and from other organs of the crocodylian gastrointestinal tract. This is also observed in mammalian Alpha- defensins (Selsted and Ouellette 2005) and has been suggested that these longer alpha-defensins evolved from these longer reptilian beta-defensin.

	Long pro-domain/mature peptide			Second Exon		
	pI	Net Charge	Mr	pI	Net Charge	Mr
PMBD25	5.09	-3	9441	9.42	7	5935
PMBD26	5.01	-3	9136	4.99	-2	6362
PMBD27	7.75	1	7491	9.18	5	4858
PMBD28	5.01	-3	9136	4.99	-2	6362
PMBD29	5.02	-3	8855	8.07	1	6236
PMBD30	5.05	-1	6458	8.66	3	4636
PMBD31	4.49	-6	8406	9.18	5	4636
PMBD32	4.51	-8	9657	9.65	8	5524
PMBD33	4.73	-6	9459	9.89	9	5370

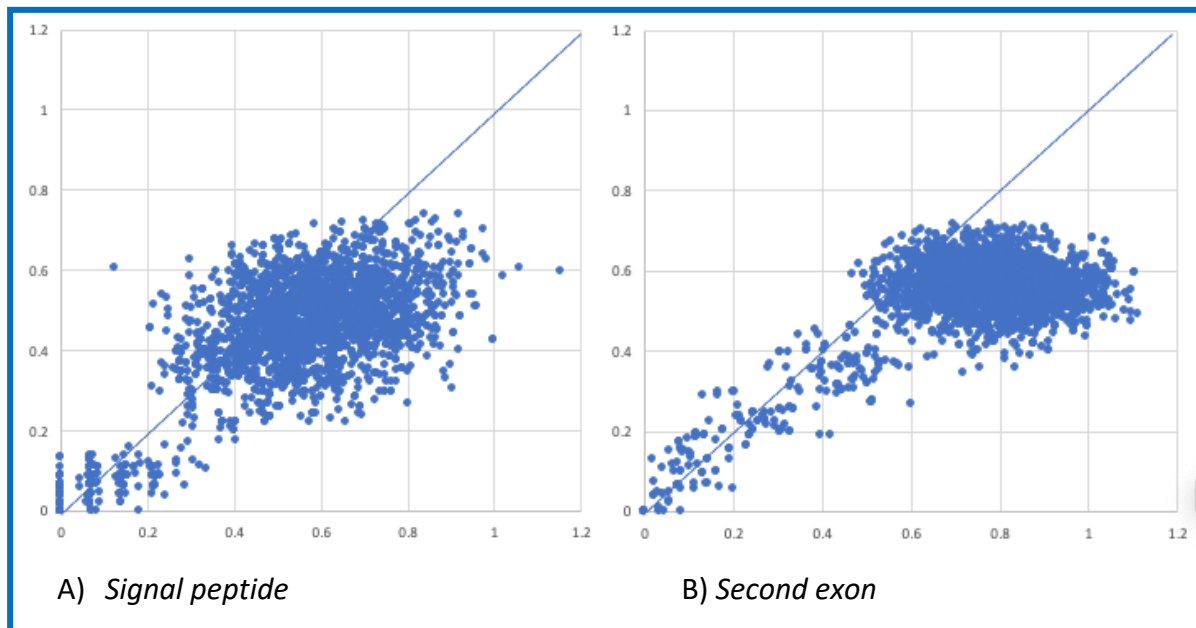
**Table 4.1 Charge differences between the longer pro-domain/mature peptides and the second exon in *P. muralis*.**

*Isoelectric point and molecular mass included.*



**Figure 4.3 Multiple sequence alignment of 80 Beta-defensin genes identified in the *P. muralis* cluster.**

Produced in Clustal X. The beta-defensin genes in this cluster show the typical conserved signal peptide at the start of the gene and in the mature peptide showing the 6 conserved cysteine residues along with glycine residues. A conserved signal peptide at the start of the genes can also be noted and varying lengths of pro peptide linking the signal peptide and the mature second exon peptide. Percentage conservation of amino acid sequences underneath. Parentheses denote closely related paralogues and intraspecific gene clustering. Signal, Pro-peptide and Mature regions are also shown by the parentheses



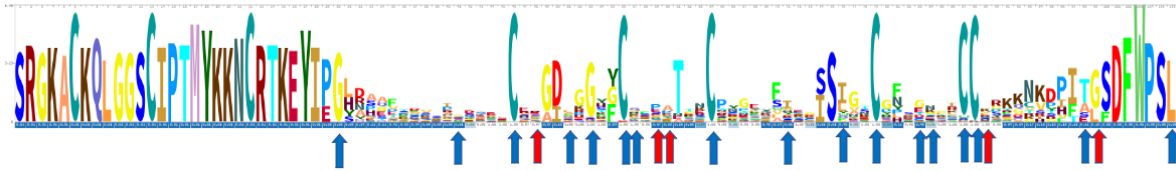
**Figure 4.4 Ratio of synonymous and nonsynonymous substitutions in *P. muralis*.**

Ratios within the signal peptide (A) and the second exon peptide (B). Graphs show nonsynonymous ( $d_N$ ) on the y axis and synonymous ( $d_S$ ) on the x axis. The diagonal lines represent  $d_N = d_S$  and is given for estimating selection pressures; dots above and below this line represent positive and purifying selection, respectively

#### 4.1.4 Selection analyses

The observable trends in the plotted proportions of synonymous and nonsynonymous substitutions within the signal peptide show more nonsynonymous substitutions than that of the second exon implying a slightly more positive selection pressure (figure 4.4). The distribution also is closer to  $d_N = d_S$ . This shows a higher degree of variation in the signal compared to the second exon. The second exon is showing more synonymous substitutions suggesting a purifying selection pressure. This could be due to the number of paralogues similar in nature, for example the genes PMBD6-23 as shown highlighted by green in the multiple sequence alignment and red in the phylogenetic analysis (figures 4.3 and 4.6).



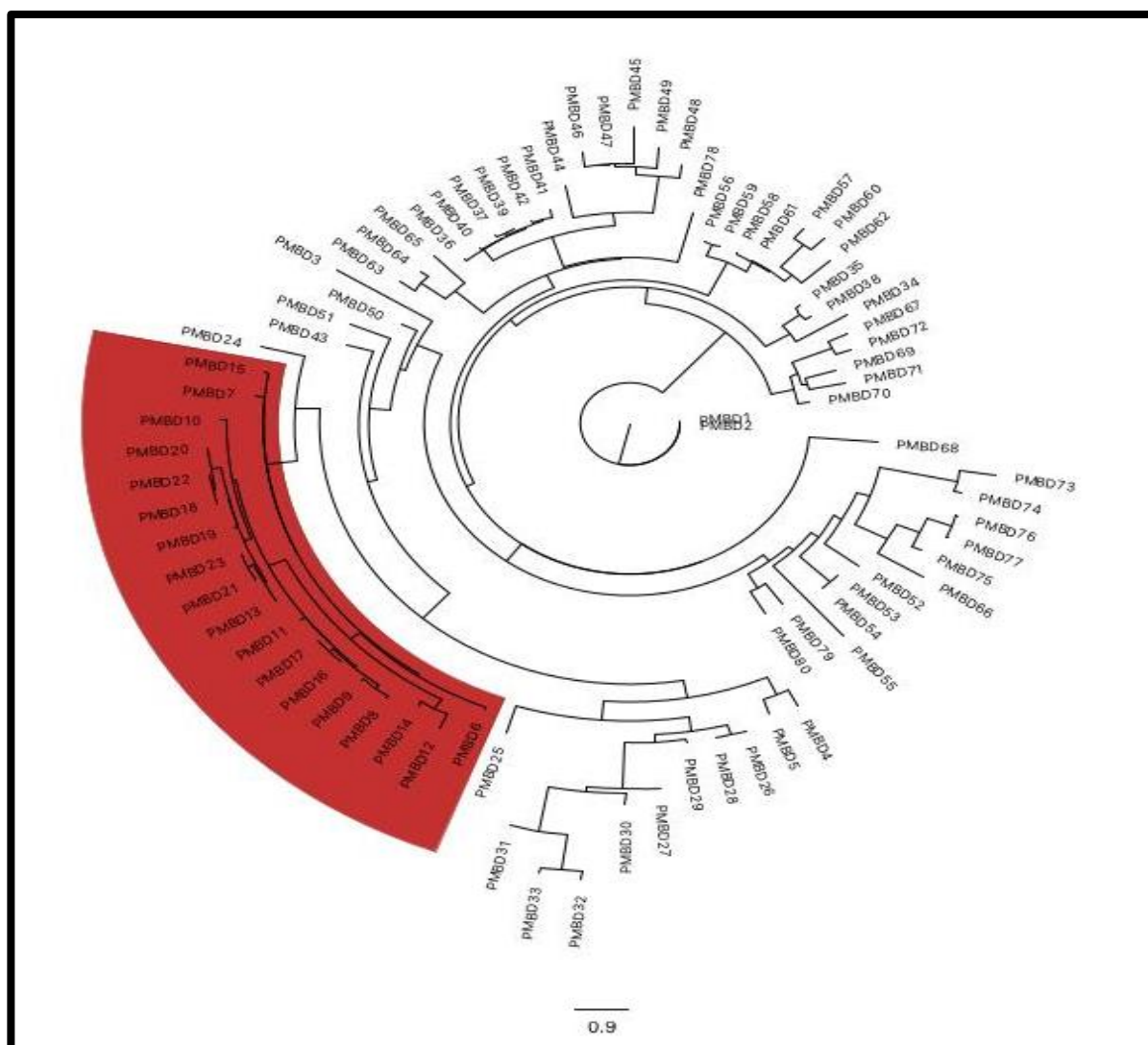


**Figure 4.5 Amino acid sequence logo of second exon of *P. muralis*.**

Sites which are undergoing positive selection (red arrow) and purifying selection (blue arrow) tested by FEL, MEME and FUBAR in HYPHY. Logo produced on Skylign.org.

Selection analysis of the individual amino acids within the peptide was performed using HyPhy (Pond *et al* 2005) via the datamonkey online server (Pond *et al* 2005a). The second exon was analysed to see which individual amino acid sites were undergoing selection. The sites that were undergoing positive or negative selection were plotted on an amino acid sequence logo produced using the Weblogo server (Crooks *et al.* 2004) (figure 4.5). There are 17 sites within the second exon region that are showing negative/purifying selection, and this is consistent with the observations in the pairwise comparison analysis, supporting previous findings (Maxwell *et al* 2003). 5 sites undergo positive selection within this peptide (red arrows in figure 4.5). The sites undergoing positive selection are within the regions which would confer a difference in the shape of the molecule whereby changing the way in which the bends in the defensin domain of the peptide form.

Phylogenetic analyses of the genes show an interspecific clustering pattern, notably PMBD6-23, which also show the presence of pseudogenes within this group (data not shown). This fits the 'birth and death' model of evolution first described by Nei and Hughes (1992). This can be seen by the phylogenetic similarities shown by the genes highlighted by red in the tree (figure 4.6). This model describes two main features: a) an interspecific gene clustering pattern and b) the presence of pseudogenes (Eirín-López *et al.* 2012).



**Figure 4.6 Phylogenetic tree of the DNA coding sequences for *P. muralis*.**

Exons 1 and 2 sequences used and produced in the IQ-tree server (Trifinopoulos et al 2016) using ultrafast analysis of 1000 bootstrap alignments. The red highlighted genes show recent duplications with high degree of conservation.

#### 4.1.5 Repeat Masking

Repeat masker was performed using query species database set to tetrapod. The *P. muralis* defensin cluster region had over all 28.47% bases masked with the predominant repeat elements being retroelements at 89.9% of the bases masked. LINES were around 73% of the retroelements and CR1 LINE being the most abundant at 44.6% of the LINES present. LTR elements accounted for about 9.6% of the retroelements. Around 4% of the repeat sequences were DNA transposons with Tc1-IS630-Pogo being the most abundant (table 4.2).

```

=====
file name: Podarcis_muralis_DNA_CHROM3_rev_comp.fa
sequences: 1
total length: 2056283 bp (2043655 bp excl N/X-runs)
GC level: 44.06 %
bases masked: 585513 bp ( 28.47 %)
=====

```

	number of elements*	length occupied	percentage of sequence
Retroelements	1506	526537 bp	25.61 %
SINEs:	471	90247 bp	4.39 %
Penelope	59	6808 bp	0.33 %
LINEs:	931	385665 bp	18.76 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	562	172406 bp	8.38 %
R1/LOA/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	11	462 bp	0.02 %
RTE/Bov-B	177	69119 bp	3.36 %
L1/CIN4	120	136753 bp	6.65 %
LTR elements:	104	50625 bp	2.46 %
BEL/Pao	1	86 bp	0.00 %
Ty1/Copia	13	11101 bp	0.54 %
Gypsy/DIRS1	30	26529 bp	1.29 %
Retroviral	42	2827 bp	0.14 %
DNA transposons	267	24083 bp	1.17 %
hobo-Activator	103	10302 bp	0.50 %
Tc1-IS630-Pogo	120	10424 bp	0.51 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	1	11 bp	0.00 %
Tourist/Harbinger	17	1511 bp	0.07 %
Other (Mirage, P-element, <u>Transib</u> )	0	0 bp	0.00 %
Rolling-circles	1	25 bp	0.00 %
Unclassified:	10	563 bp	0.03 %
Total interspersed repeats:		551183 bp	26.80 %
Small RNA:	18	982 bp	0.05 %
Satellites:	1	48 bp	0.00 %
Simple repeats:	721	30309 bp	1.47 %
Low complexity:	67	3298 bp	0.16 %

\* most repeats fragmented by insertions or deletions  
have been counted as one element

The query species was assumed to be tetrapods  
RepeatMasker version 4.1.2-p1 , default mode

run with rmblastn version 2.2.27+  
FamDB: CONS-Dfam\_withRBRM\_3.3

**Table 4.2 Repeat masker summary for *P. muralis*.**

Displaying the different repeat sequences within the *P. muralis* cluster region. The tetrapod database was used as a reference for the repeat sequence matches in RepeatMasker program.

## 4.2 *Lacerta agilis* – Sand Lizard



### ***Distribution map and image of L. Agilis***

Photos obtained - <https://www.eurolizards.com/lizards/lacerta-agilis/>

The Sand Lizard is a lacertid lizard which is distributed across most of Europe and there are several subspecies. They are also a sexually dimorphic legged lizard. In Northwest Europe both sexes show lateral and dorsal strips of ocellated (eye-shaped) markings. They can grow up to 20cm long and can live up to 20 years. The males turn bright green during the mating season and fading during late summer. Their Habitat is largely restricted to lowland heathland and sandunes, hence the name of the lizard. They feed on fruit and flower heads as well as insects, slugs and spiders. The females make their burrow in the sand in which they lay up to 14 eggs which hatch in the late summer.

### **4.2.1 Data Mining and Cluster assembly**

The genomic sequence data was accessed through the NCBI genome assembly database. This species was again, chosen as the assembly was at chromosomal level therefore it would likely have the Beta-defensin cluster region intact allowing a full in-depth search of the beta-defensins present in the area. The GenBank assembly accession number is GCF\_009819535.1 and was submitted to the database on 31/12/2019. Using CTSB as the reference for the start of the analysis, the chicken CTSB amino acid sequence was once again used as a query to search against the genome using the tBLASTn program. XPO1 was expected to be the other

flanking gene of the cluster, and this was also searched for in the genome. The cluster region was then established and further probed for to search for the beta-defensins present.

Using the concatemer approach, amino acid sequences from the those obtained in the previous analysis in *P. muralis* were utilised to construct the query sequence in which this region was searched using the tBLASTn program on the NCBI server. The region found was masked for repeat sequences using Repeat Masker program and then this was translated into a 6-frame translation using the EMBOSS sixpack program on the EMBL-EBI server. This output file was used to highlight matches from the tBLASTn concatemer query searches.

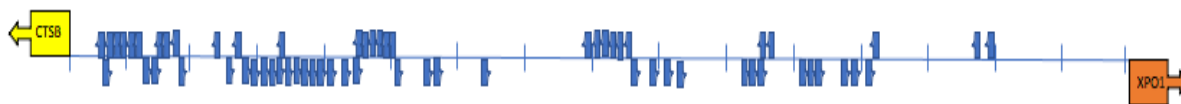
Further iterative searches were performed with the *Chrysemys picta bellii* DNA sequences produced during the method development and queried against the region using the BLASTn program. However, there were still suspected exons missing from the construct and therefore gene finding programs as outlined previously were employed. Finally, regions of more than 3000bp of the repeats determined in the repeat masker analysis but not in the vicinity of already resolved exons and downstream from the poly adenylation signal can then be searched to exclude all potential regions where Beta-defensin exons may reside.

Splice site prediction was finalised using the online server by the Berkeley Drosophila Genome Project (Reese *et al.* 1997) and amino translations were ascertained from the DNA sequences of potential exons. Finally, iterative searches were performed using the newly identified beta-defensins against the cluster region.

#### **4.2.2 Cluster organisation and Beta-defensin sequences**

The cluster was identified and was syntenic with the *P. muralis* cluster and the Beta-defensin genes were found between CTSB and XPO1. A total of 64 genes were identified (appendix 1.4) spanning a region of approximately 1.514 Mb long (figure 4.7). The cluster was found on chromosome 3 (Genbank sequence NC\_046314.1) between locations 110816535-112330572 for which the reverse complement was used for further analysis so to start the cluster at the CTSB gene. The naming of the genes used a prefix to the order number and uses an abbreviation of the species name, in this case LABD.





**Figure 4.7 Genomic organisation of the *L. agilis* Beta-defensin cluster.**

Each vertical line represents 100kbp along the chromosome. The blue blocks represent the full genes, and the arrows dictate their orientation. The size of the squares (genes) or the distance between them are only representative and are not proportional to their actual size in the genome.

The beta-defensins that were discovered in this analysis show the typical defensin structure in that the genes consist of two exons except for LABD11 and 12 where a potential third exon was found. This will need to be confirmed with laboratory analyses. The first exon encodes a signal peptide followed by the second exon encoding the mature peptide that consists of the defensin motif with common 6 cysteine domain and a glycine located in the N-terminal region, two positions upstream from the second cysteine, while another glycine is placed two to four positions upstream from the fourth cysteine.

#### **4.2.3 Physical Properties**

As with the findings in *P. muralis* each beta-defensin gene identified in this genome possesses a conserved signal peptide, and this is seen in the sequence analyses using SignalIP – 5.0 server (Almagro Armenteros *et al* 2019) (appendix 1.6) There is a wide range of charges from -11 to +11 with some of the beta-defensins in this cluster being highly anionic which is contrary to the notion that beta-defensins are usually cationic. One such defensin, LABD21 (appendix 1.5), has a charge of -11 but this is similar to what has been reported in crocodylians (Tang *et al* 2018) in that it has a long anionic pro-domain, and this has been described (Michaelson *et al* 1992) as a mechanism in which the pro-domain counterbalances the cationic charge of the active Beta-defensin during synthesis. Table 4.3 Shows the charges between the long pro-domain beta-defensins minus the signal sequences and the charge of the second exon, which may closely represent the mature active form. From this there is a difference in charge supporting the observations made by Michaelson *et al* (1992). Tang *et al* (2018) also states that this expression pattern has been observed in the small intestine and

from other organs of the crocodylian gastrointestinal tract. This is also observed in mammalian alpha-defensins (Selsted and Ouellette 2005) and has been suggested that these longer alpha-defensins evolved from these longer reptilian beta-defensins.

	Long pro-domain/mature peptide			Second Exon		
	pI	Net Charge	Mr	pI	Net Charge	Mr
LABD15	5.08	-4	9697	9.55	7	5865
LABD16	8.62	3	9152	9.55	7	6046
LABD17	5.26	-2	7932	6.98	0	6223
LABD18	4.82	-5	9130	4.99	-2	6362
LABD19	5.63	-1	6506	8.68	3	4715
LABD20	5.73	-1	8142	8.7	3	4885
LABD21	4.1	-11	9242	7.75	1	5043
LABD22	4.57	-7	9547	9.4	7	5459
LABD23	4.68	-1	9455	9.49	7	5367

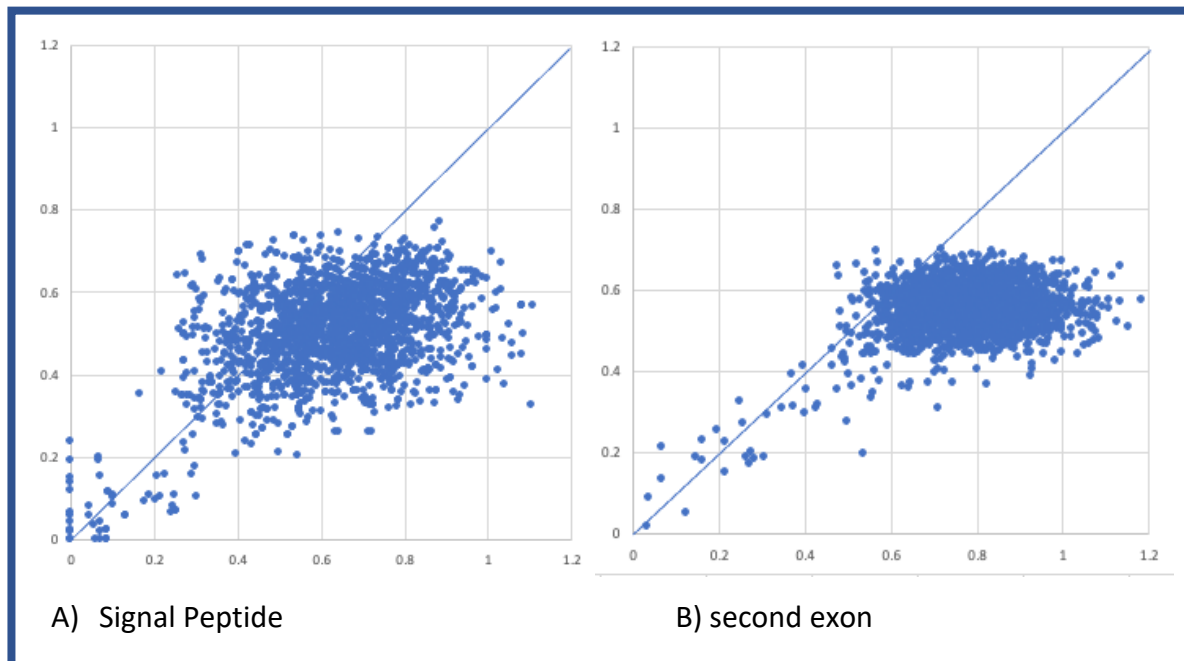
**Table 4.3 Charge differences between the longer pro-domain/mature peptides and the second exon in *L. agilis*.**

*Isoelectric point and molecular mass included.*

#### **4.2.4 Selection analyses**

Evolutionary analyses were conducted on the beta-defensin genes in the cluster. Given that beta-defensin clusters arise from gene duplication and paralogous to each other, all possible pairwise comparisons of the genes within the cluster were used to investigate the proportion of non-synonymous ( $d_N$ ) to synonymous ( $d_S$ ) substitutions. This was done separately for the signal peptide and the mature peptide (Figure 4.8). The distributions for the signal peptide show that there is a slight distribution towards positive selection, but the overall look of the data suggest that this undergoing neutral selection. The second exon shows a greater degree of purifying selection compared to the signal peptide. It could be hypothesised that this is due to the number of highly similar paralogues within the cluster region. However, the

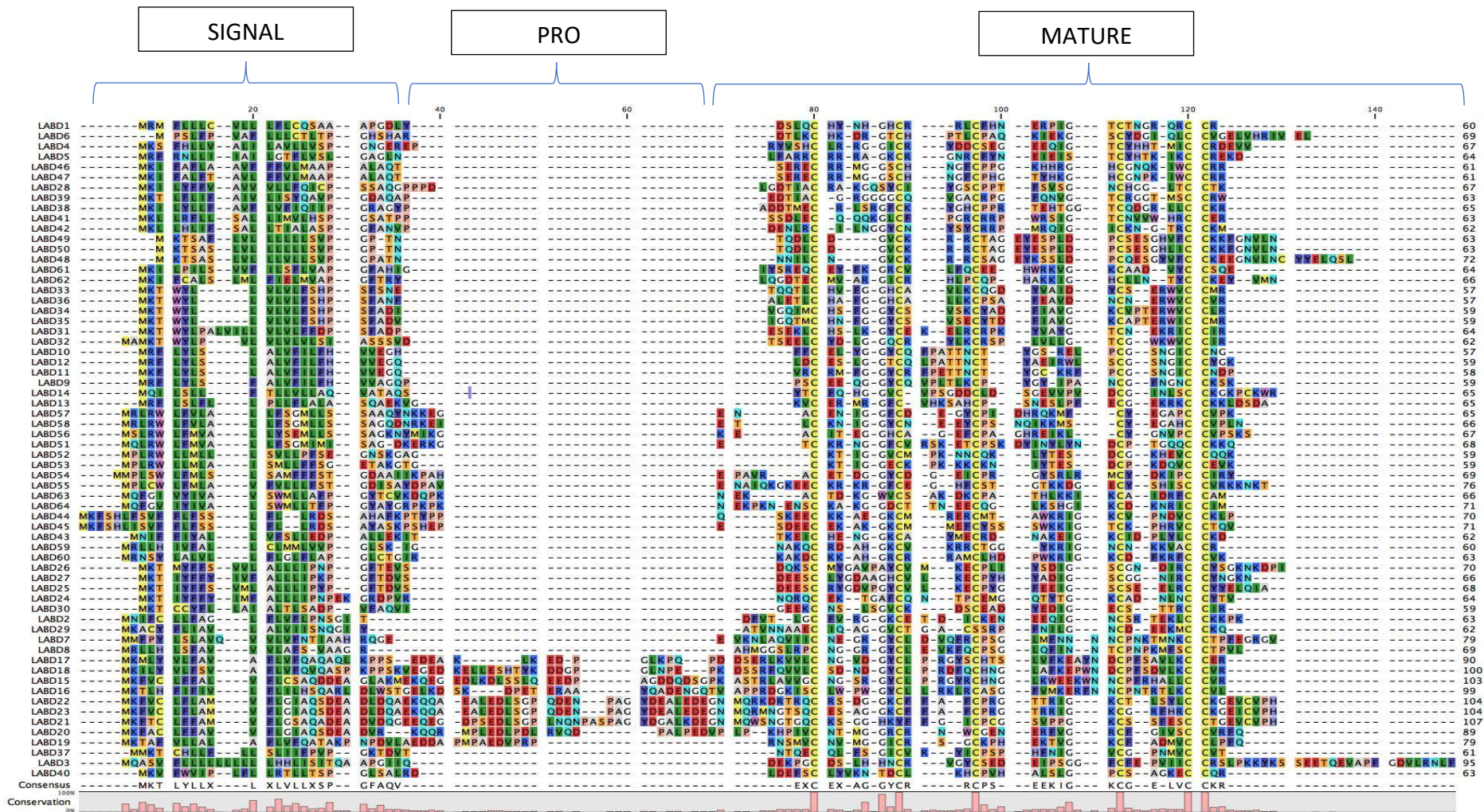
multiple sequence alignment shows more conservation in the similarity percentages (Figure 4.9).



**Figure 4.8 Proportions of synonymous and nonsynonymous substitutions in *L. agilis*.**

Ratios within the signal peptide (A) and the second exon peptide (B) show synonymous ( $d_N$ ) on the x axis and nonsynonymous ( $d_S$ ) on the y axis. The diagonal lines represent  $d_N = d_S$  and is given for estimating selection pressures; dots above and below this line represent positive and purifying selection, respectively.

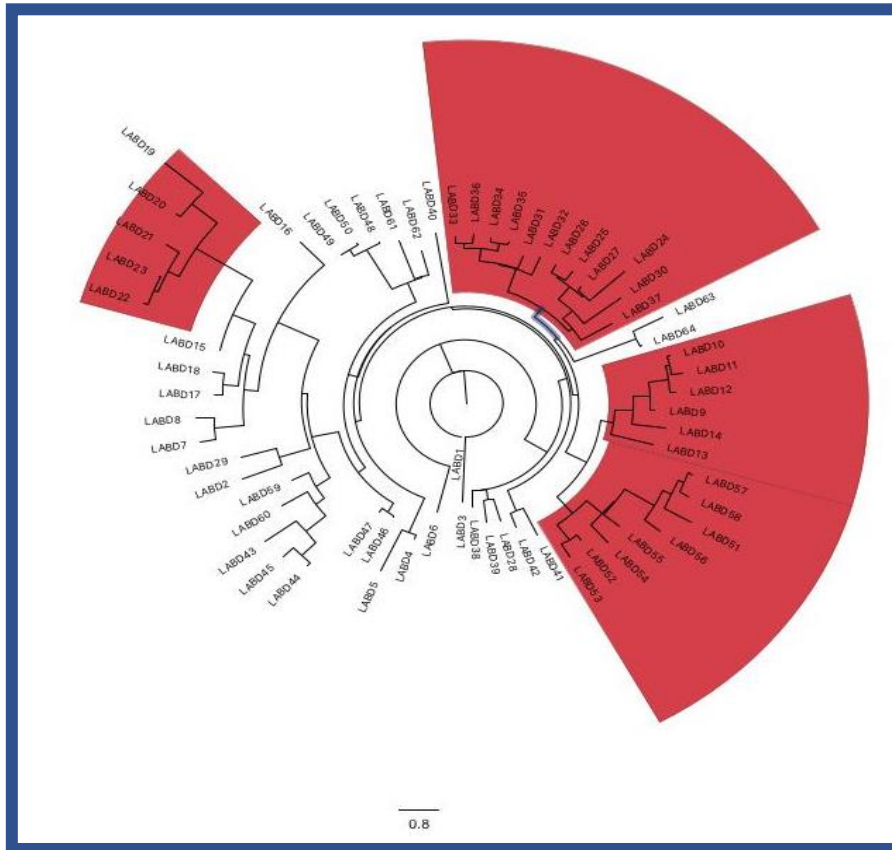
Selection analysis of the individual amino acids within the peptide was performed using HyPhy (Pond *et al* 2005) via the datamonkey online server (Pond *et al* 2005a). The sites that were undergoing positive or negative selection were plotted on an amino acid sequence logo produced using the Weblogo server (Crooks *et al.* 2004). There were 21 sites undergoing purifying selection and only two sites showing positive selection. This backs up the analysis done looking into the  $d_N/d_S$ . These indicate that beta-defensins identified in lizards could be more stable. It also suggests that the second exon may not have had the pressure conditions to diversify.



**Figure 4.9 Multiple sequence alignment of 64 Beta-defensin genes identified in the *L. agilis* cluster.** Produced in Clustal X. The beta-defensin genes in this cluster show the typical conserved signal peptide at the start of the gene and in the mature peptide showing the 6 conserved cysteine residues along with glycine residues. A conserved signal peptide at the start of the genes can also be noted and varying lengths of pro peptide linking the signal peptide and the mature second exon peptide. Percentage conservation of amino acid sequences underneath. Signal, Pro-peptide and Mature regions are also shown by the parentheses.







**Figure 4.11** Phylogenetic tree of the DNA coding sequences of *L. agilis*.

Exons 1 and 2 were used to produce tree in the IQ-tree server (Trifinopoulos et al 2016) using ultrafast analysis of 1000 bootstrap alignments. The red highlighted genes show recent duplications with high degree of conservation.

#### 4.2.5 Repeat sequence landscape

Repeat masker was performed using query species database set to tetrapod. The *L. agilis* defensin cluster region had over all 26.10% bases masked with the predominant repeat elements being retroelements at 93.7% of bases masked. LINES were around 73.3% of the retroelements and CR1 LINE being the most abundant at 75.5% of the LINES present. LTR elements accounted for about 7.2% of the retroelements. Around 5.3% of the repeat sequences were DNA transposons with Hobo-Activator being the most abundant (table 4.4).

```

=====
file name: lacerta_agilis_CTSB-XP0_rev_comp.fa
sequences: 1
total length: 1514039 bp (1513939 bp excl N/X-runs)
GC level: 43.98 %
bases masked: 395214 bp ( 26.10 %)
=====

```

	number of elements*	length occupied	percentage of sequence
Retroelements	1366	340376 bp	22.48 %
SINES:	362	58829 bp	3.89 %
Penelope	157	15750 bp	1.04 %
LINES:	936	257066 bp	16.98 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	489	121684 bp	8.04 %
R1/L0A/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	7	356 bp	0.02 %
RTE/Bov-B	215	87514 bp	5.78 %
L1/CIN4	69	31772 bp	2.10 %
LTR elements:	68	24481 bp	1.62 %
BEL/Pao	1	196 bp	0.01 %
Ty1/Copia	5	2214 bp	0.15 %
Gypsy/DIRS1	18	17392 bp	1.15 %
Retroviral	37	2741 bp	0.18 %
DNA transposons	233	21019 bp	1.39 %
hobo-Activator	111	9551 bp	0.63 %
Tc1-IS630-Pogo	61	7380 bp	0.49 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	25	2328 bp	0.15 %
Other (Mirage, P-element, <u>Transib</u> )	0	0 bp	0.00 %
Rolling-circles	7	113 bp	0.01 %
Unclassified:	5	424 bp	0.03 %
Total interspersed repeats:		361819 bp	23.90 %
Small RNA:	21	1332 bp	0.09 %
Satellites:	2	25 bp	0.00 %
Simple repeats:	722	29473 bp	1.95 %
Low complexity:	53	2706 bp	0.18 %

```

=====
* most repeats fragmented by insertions or deletions
  have been counted as one element

The query species was assumed to be tetrapods
RepeatMasker version 4.1.2-p1 , default mode

run with rmblastn version 2.2.27+
FamDB: CONS-Dfam_withRBRM_3.3
=====

```

**Table 4.4 Repeat masker summary in *L. agilis*.**

*Displaying the different repeat sequences within the L. agilis cluster region. The tetrapod database was used as a reference for the repeat sequence matches in RepeatMasker program.*

### 4.3 *Zootoca vivipara* – Viviparous Lizard/Common Lizard



**Distribution Map and image of *Z. vivipara***

Photos obtained - <https://www.eurolizards.com/lizards/zootoca-vivipara/>

The Viviparous/Common lizard is the only lacertid viviparous species, meaning that it does not lay eggs and consequently can tolerate colder climates. It also belongs to the monotypic genus *Zootoca*. Its distribution covers large areas of northern Eurasia and to Japan and even has populations in the arctic circle. It could possibly be the most successful living reptile. Their colouration is typically brown but grey and olive colours have also been observed. Males have brightly coloured undersides; however, females show a greater degree of polymorphism with colours of yellow, orange or a mixture of the two. They can grow up to 12cm excluding the tail and the tail can be twice the length of the body. They mate in April/May and females give birth to 3-10 young after 3 months. The diet is mainly invertebrates, mostly insect or spiders.

#### 4.3.1 Data Mining and Cluster assembly

The genomic sequences data was accessed through the NCBI genome assembly database. This species was again, chosen as the assembly was at chromosomal level therefore it would likely have the beta-defensin cluster region intact allowing a full in-depth search of the beta-defensins present in the area. The genome assembly (GenBank assembly accession GCF\_011800845) and was submitted to the database on 01/04/2020. Using CTSB as the reference for the start of the analysis, the chicken CTSB amino acid sequence was once again used as a query to search against the genome using the tBLASTn program. As before XPO1



was expected to be the other flanking gene of the cluster and the was also searched for in the genome. The cluster region was then established and further probed for to search for the beta-defensins present.

Using the concatemer approach, amino acid sequences from those obtained in the previous *P. muralis* and *L. agilis* analysis were utilised to construct the query sequence in which this region was searched using the tBLASTn program on the NCBI server. The region found was masked for repeat sequences using RepeatMasker program and then this was translated into a 6-frame translation using the EMBOSS sixpack program on the EMBL-EBI server. This output file was used to highlight matches from the tBLASTn concatemer query searches.

Further searches were performed with the DNA sequences produced from the *P. muralis* and *L. agilis* analysis and queried against the region using the BLASTn program. However, there were still suspected exons missing from the construct and therefore gene finding programs as outlined previously were employed. Finally, regions of more than 3000bp of the repeats determined in the repeat masker analysis but not in the vicinity of already resolved exons and downstream from the poly adenylation signal can then be searched to exclude all potential regions where Beta-defensin exons may reside.

Splice site prediction was finalised using the online server by the Berkeley Drosophila Genome Project (Reese *et al.* 1997) and amino translations were ascertained from the DNA sequences of potential exons. Finally, iterative searches were performed using the newly identified beta-defensins against the cluster region.

#### **4.3.2 Cluster organisation and Beta-defensin sequences**

The cluster was identified and was syntenic with the snake clusters in that the beta-defensin genes were discovered to reside between CTSB and XPO1. A total of 34 genes were identified (appendix 1.7) spanning a region of approximately 968 kbps long (figure 4.12). The cluster was found on Genbank sequence NC\_048607.1 between locations 5470382-6439240 and the start of the cluster is at the last codon position of the CTSB gene. The naming of the genes used a prefix to the order number and was an abbreviation of the species name, in this case ZVBD.



**Figure 4.12 Genomic organisation of the *Z. vivipara* Beta-defensin cluster.**

Each vertical line represents 100kbp along the chromosome. The blue blocks represent the full genes, and the arrows dictate their orientation. The size of the squares (genes) or the distance between them are only representative and are not proportional to their actual size in the genome.

### 4.3.3 Physical Properties

The beta-defensins that were discovered in this analysis show the classical structure showing that the genes consist of two exons except for ZVBD17, 18 and 19 where a suspected third exon was found. This, however, needs to be confirmed with laboratory analyses. The first encodes a signal peptide followed by the second exon encoding the mature peptide that consists of the typical defensin motif with common 6 cysteine domain and a glycine located in the N-terminal region, two positions upstream from the second cysteine, while another glycine is placed two or three positions upstream from the fourth cysteine.

As with the findings in *P. muralis* and *L. agilis* each beta-defensin gene identified in this genome possesses a conserved signal peptide, and this is seen in the sequence analyses using SignalIP – 5.0 server (Almagro Armenteros *et al* 2019) (appendix 1.9). There is a wide range of charges and some of the beta-defensins in this cluster are highly anionic which is contrary to the notion that beta-defensins are usually cationic. One such defensin, LABD20 (appendix 1.8), has a charge of -9 concurring with crocodilian findings, (Tang *et al* 2018) in that it has a long anionic pro-domain, and this has been described (Michaelson *et al* 1992) as a mechanism in which the pro-domain counterbalances the cationic charge of the active Beta-defensin during synthesis. Table 4.5 shows the charges between the long pro-domain beta-defensins minus the signal sequences and then the charge of the 2<sup>nd</sup> exon, which may closely represent the mature active form. From this there is a difference in charge supporting the observations made by Michaelson *et al* (1992). Tang *et al.* (2018) also states that this expression pattern has been observed in the small intestine and from other organs of the crocodilian gastrointestinal tract. This is also observed in mammalian alpha-defensins

(Selsted and Ouellette 2005) and has been suggested that these longer alpha-defensins evolved from these longer reptilian beta-defensin.

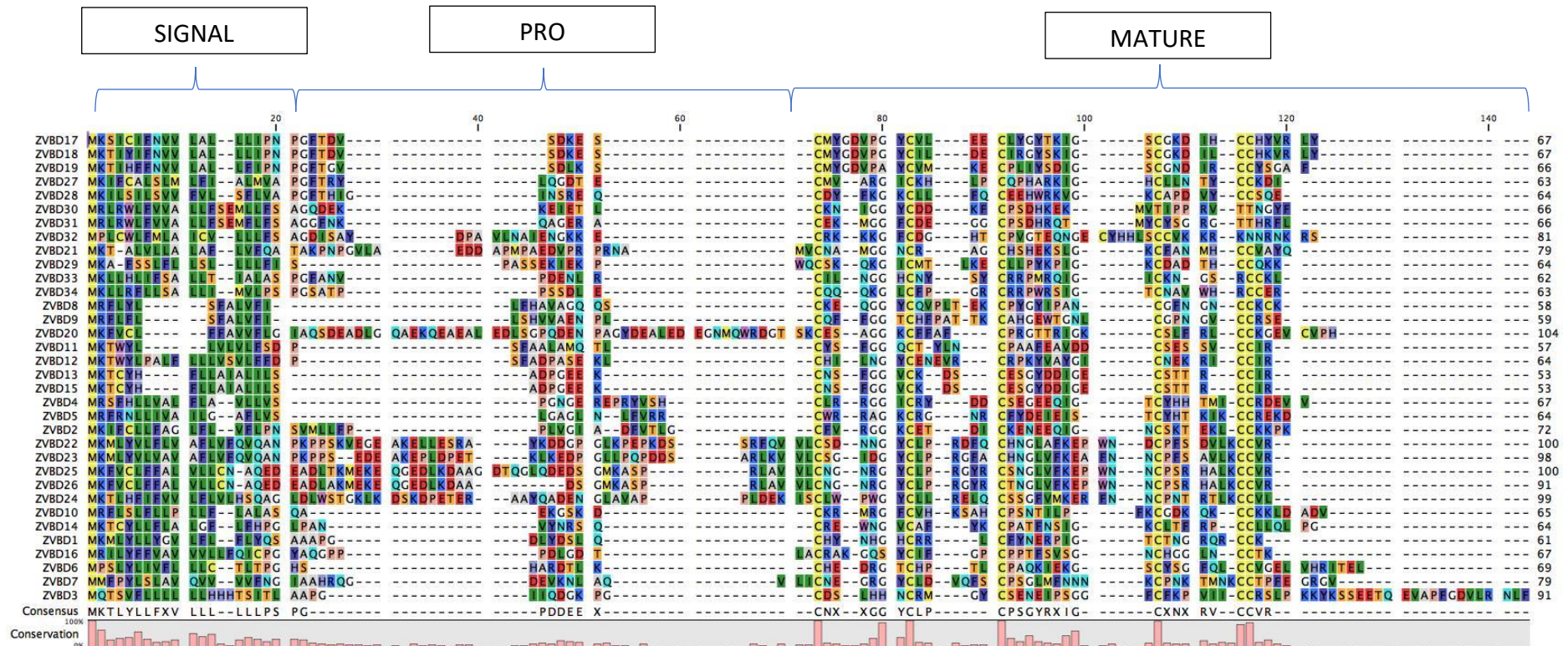
	Long pro-domain/mature peptide			Second Exon		
	pI	Net Charge	Mr	pI	Net Charge	Mr
ZVBD20	4.41	-9	9348	8.26	2	5860
ZVBD21	6.97	0	6401	8.68	3	4668
ZVBD22	5.77	-1	9115	6.98	0	6375
ZVBD23	5.26	-2	8629	8.46	2	6041
ZVBD24	5.31	-1	9025	8.33	2	6019
ZVBD25	5.33	-2	9318	9.77	8	5783
ZVBD26	7.76	1	8358	9.77	8	5797

**Table 4.5 Charge differences between the longer pro-domain/mature peptides and the second exon in *Z. vivipara*.**

*Predicted Isoelectric point and molecular mass included.*

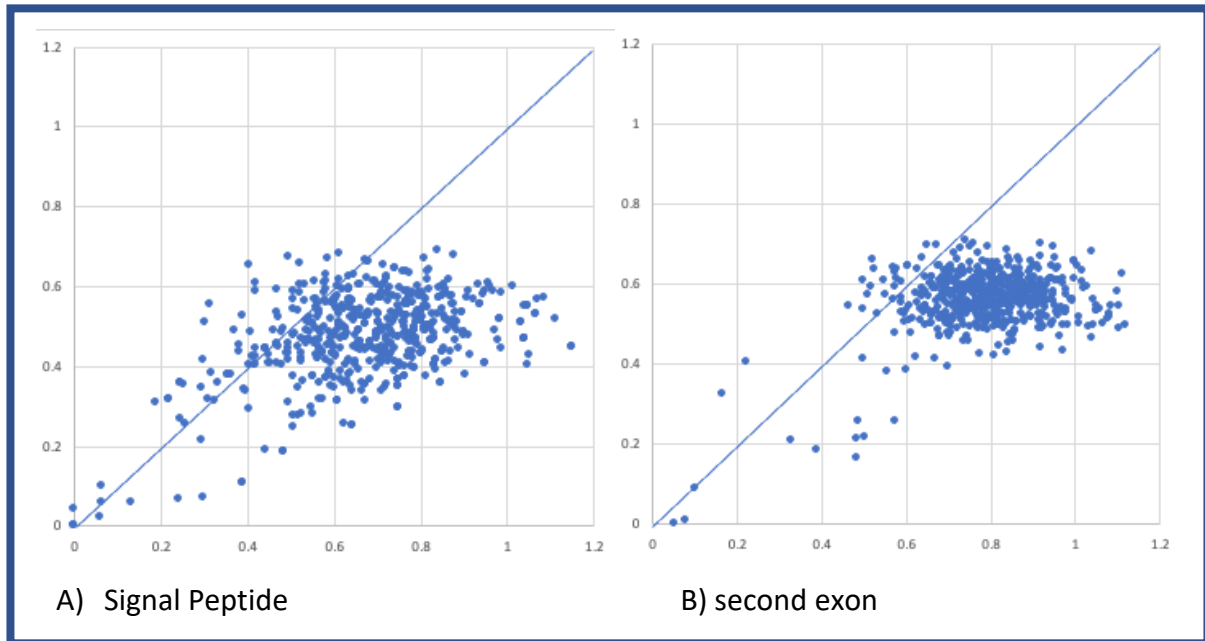
#### **4.3.4 Selection analyses**

Evolutionary analyses were conducted on the beta-defensin genes in the cluster. Given that beta-defensin clusters arise from gene duplication and paralogous to each other, pairwise comparisons of each gene against all combinations were performed. The proportions of non-synonymous ( $d_N$ ) to synonymous ( $d_S$ ) substitutions were determined and plotted against each other. This was done separately for the signal peptide and the mature peptide (figure 4.14). Multiple sequence alignments were produced in CLUSTALX (Larkin *et al* 2007) and Codon alignments from this were made using the PAL2NAL server (Suyama *et al* 2006). This codon alignment was used in the SNAP program on the HIV database server ([https://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html?sample\\_input=1](https://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html?sample_input=1))



**Figure 4.13 Multiple sequence alignment of 34 Beta-defensin genes identified in the *Z. vivipara* cluster.**

Produced in Clustal X. the beta-defensin genes in this cluster show the typical conserved signal peptide at the start of the gene and in the mature peptide showing the 6 conserved cysteine residues along with glycine residues. A conserved signal peptide at the start of the genes can also be noted and varying lengths of pro peptide linking the signal peptide and the mature second exon peptide. Percentage conservation of amino acid sequences underneath. Signal, Pro-peptide and Mature regions are also shown by the parentheses.

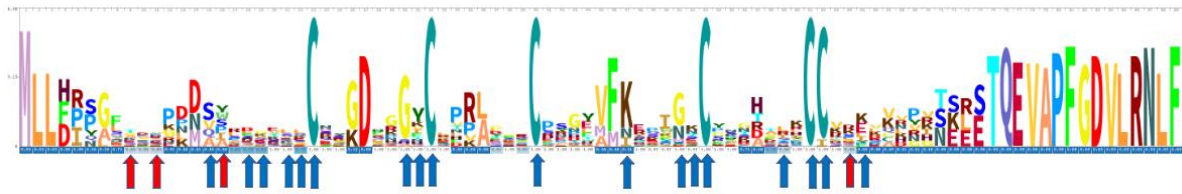


**Figure 4.14 Ratio of synonymous and nonsynonymous substitutions in *Z. vivipara*.**

Ratios within the signal peptide (A) and the second exon peptide (B) are shown with synonymous ( $d_N$ ) on the x axis and nonsynonymous ( $d_S$ ) on the y axis. The diagonal lines represent  $d_N = d_S$  and is given for estimating selection pressures; dots above and below this line represent positive and purifying selection, respectively.

As with the *P. muralis* many of the genes within the signal peptide and second exon showed that the pairwise comparisons exhibited either neutral or purifying selection with a few undergoing positive selections. The distributions were similar when the data points were plotted.

Selection analysis of the individual amino acids within the peptide was performed using HyPhy (Pond *et al* 2005; Pond *et al* 2005a). The second exon was analysed to see which individual amino acid sites were undergoing selection. The sites that were undergoing positive or negative selection were plotted on an amino acid sequence logo produced using the Weblogo server (Crooks *et al.* 2004).



**Figure 4.15 Amino acid sequence logo of *Z. vivipara* second exons.**

Sites which are undergoing positive selection (red arrow) by one of more tests and purifying selection (blue arrow) tested by FEL, MEME and FUBAR in HYPHY. Logo produced on Skyline.org.

The analysis looking at the selection of different sites shows that there are 18 sites undergoing negative/purifying selection through the beta-defensin mature domain with just four sites at the beginning and the end showing a positive selection. This backs up the analysis done looking into the  $d_N/d_S$ .

### **3.4.5 Repeat Sequence Landscape**

Repeat masker was performed using query species database set to tetrapod. The *Z. vivipara* defensin cluster region had over all 26.10% bases masked with the predominant repeat elements being retroelements at 85% of bases masked. LINES were around 82.3% of the retroelements and CR1 LINE being the most abundant at 55.4% of the LINES present. LTR elements accounted for about 2.7% of the retroelements. Around 7% of the repeat sequences were DNA transposons with Hobo-Activator being the most abundant (table 4.6).



```

=====
file name: Zootoca_vivipara_contig.fa
sequences: 1
total length: 968859 bp (922956 bp excl N/X-runs)
GC level: 44.39 %
bases masked: 252885 bp ( 26.10 %)
=====

```

	number of elements*	length occupied	percentage of sequence
Retroelements	843	215037 bp	22.19 %
SINEs:	204	32044 bp	3.31 %
Penelope	89	10876 bp	1.12 %
LINEs:	614	177149 bp	18.28 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	342	98157 bp	10.13 %
R1/LOA/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	5	171 bp	0.02 %
RTE/Bov-B	132	49467 bp	5.11 %
L1/CIN4	46	18478 bp	1.91 %
LTR elements:	25	5844 bp	0.60 %
BEL/Pao	0	0 bp	0.00 %
Ty1/Copia	1	730 bp	0.08 %
Gypsy/DIRS1	7	3753 bp	0.39 %
Retroviral	16	982 bp	0.10 %
DNA transposons	195	17908 bp	1.85 %
hobo-Activator	84	7623 bp	0.79 %
Tc1-IS630-Pogo	61	5176 bp	0.53 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	30	3976 bp	0.41 %
Other (Mirage, P-element, Transib)	0	0 bp	0.00 %
Rolling-circles	6	507 bp	0.05 %
Unclassified:	4	265 bp	0.03 %
Total interspersed repeats:		233210 bp	24.07 %
Small RNA:	21	905 bp	0.09 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	380	15951 bp	1.65 %
Low complexity:	51	2596 bp	0.27 %

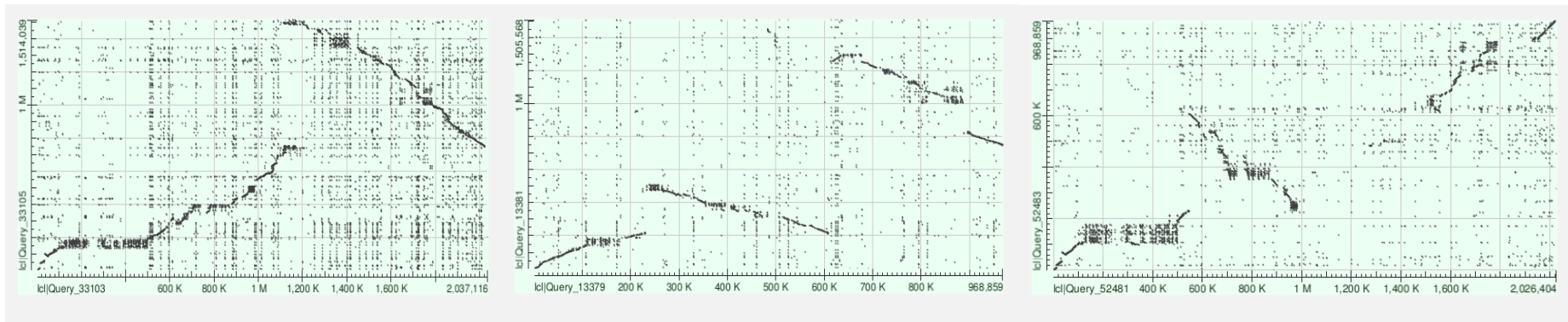
\* most repeats fragmented by insertions or deletions have been counted as one element

The query species was assumed to be tetrapods  
RepeatMasker version 4.1.2-p1 , default mode

run with rmblastn version 2.2.27+  
FamDB: CONS-Dfam\_withRBRM\_3.3

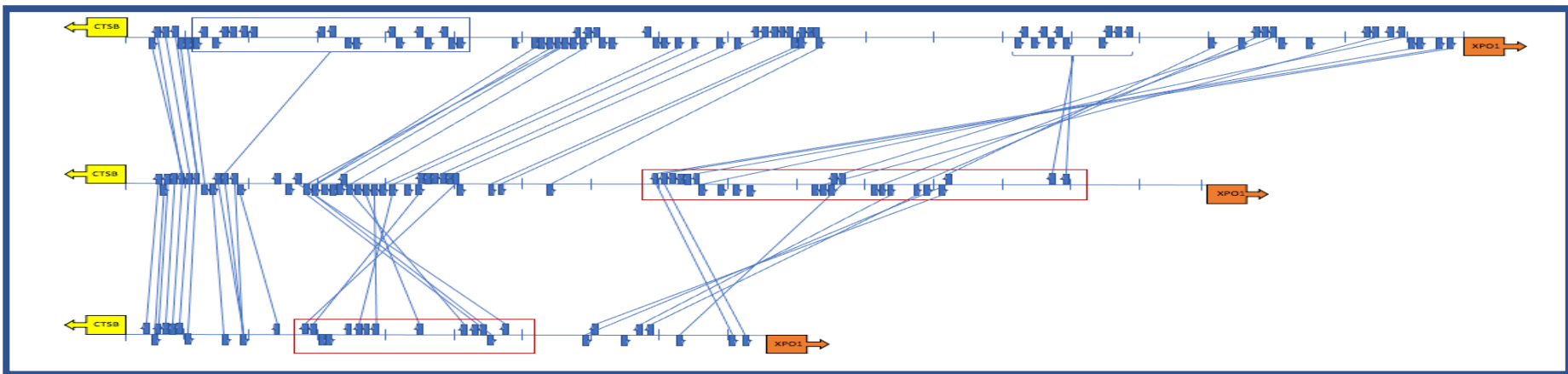
**Table 4.6 Repeat masker summary in *Z. vivipara*.**

Displaying the different repeat sequences within the *Z. vivipara* cluster region. The tetrapod database was used as a reference for the repeat sequence matches in RepeatMasker program



A) *P. muralis* (x-axis) -vs- *L. agilis* (y-axis)    B) *Z. vivipara* (x-axis) -vs- *L. agilis* (y-axis)    C) *P. muralis* (x-axis) -vs- *Z. vivipara* (y-axis)

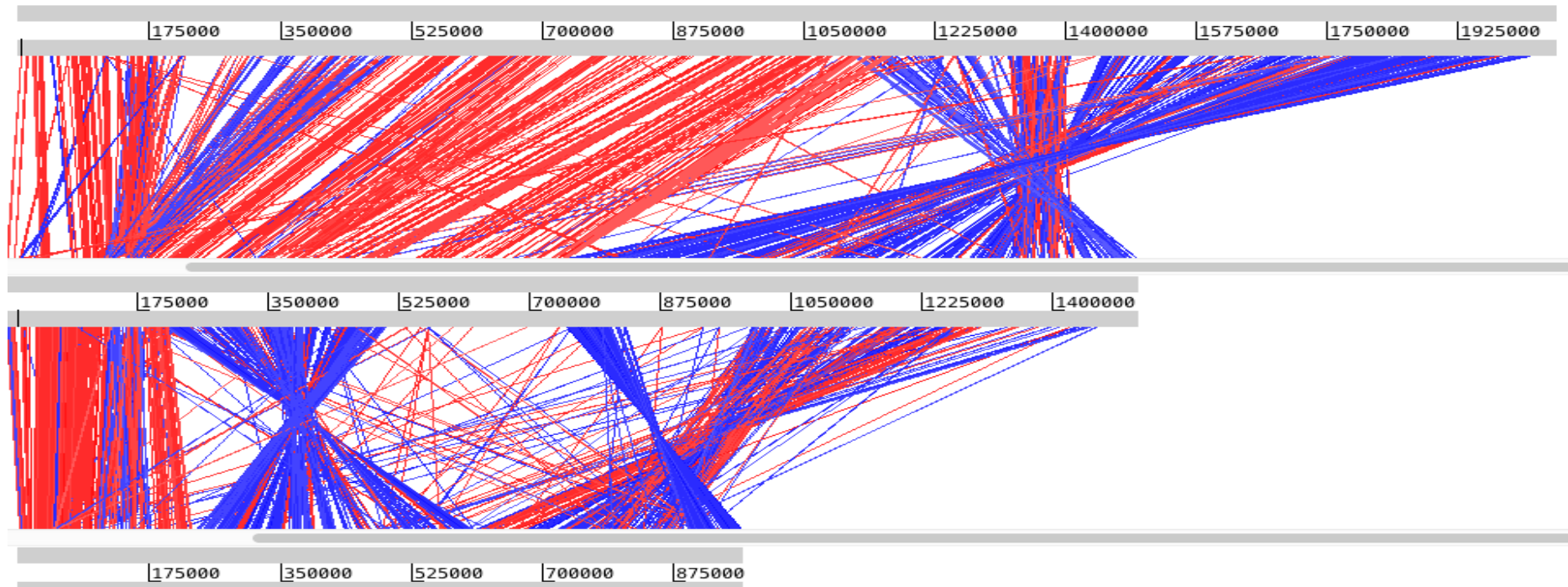
**Figure 4.16 Dot plots of Lizard species. Masked cluster genomic sequences plotted against another showing regions that have been inverted.**



**Figure 4.17 Synteny between Lizard clusters.**

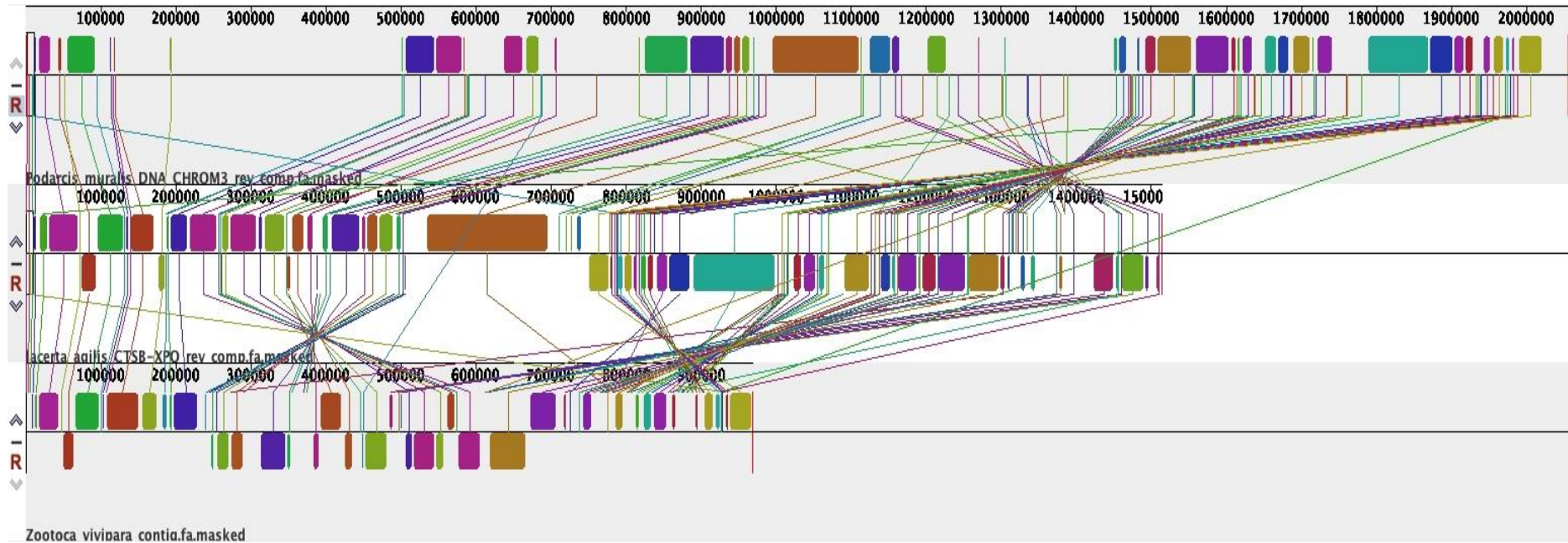
Top cluster is *P. muralis*, middle is *L. agilis* and bottom is *Z. vivipara*. Genes are represented by dark blue boxes on the scale line. Blue lines joining paralogous genes were identified by phylogenetic analysis. Blue box represents species specific orthologous genes of *P. muralis* that have duplicated in that region. Red Boxes show genomic regions that have inverted in relation to *P. muralis* as the reference cluster.





**Figure 4.18 Cluster alignment of each cluster region in Lizards.**

*Produced by sequences matches produced by BLAST and visualised in Artemis ACT. Top cluster is P. muralis, middle is L. agilis and bottom is Z. vivipara. Red connecting line are running in the same orientation and blue connecting lines in reverse orientation. Length in bp is shown by the grey scale bars.*



**Figure 4.19 Cluster alignment of each cluster region produced by sequences matches produced by MAUVE sequence aligner.**

Top cluster is *P. muralis*, middle is *L. agilis* and bottom is *Z. vivipara*. Coloured blocks show sequence regions of high similarity and connecting lines join these regions. Coloured blocks above the line show sequences running in forward direction and below in the reverse direction.

#### **4.5 Conservation of synteny and genomic reorganisation**

Comparison dot plots were produced with the masked DNA sequences of the cluster regions. There is a high degree of homology within the sequences and the cluster regions, however, some genomic inversions have occurred. All the cluster regions are flanked by CTSB and XPO1. The genes leading from CTSB show a high level of similarity in all the cluster regions. This could indicate that these genes arose before the evolution and separation into the separate species we see now.

The blue box on the gene map (figure 4.17) represents a large region of duplication present in the *P. muralis* cluster but absent in both the *L. agilis* and *Z. vivipara* clusters. This can be seen as an absence on the alignment in (figure 4.19). The blue parenthesis on the *P. muralis* cluster also indicates another region that has undergone a series of duplications which are absent on the *L. agilis* and *Z. vivipara* clusters. The genes are orthologous to single gene in the other clusters. Using the *P. muralis* as a reference genomic region highlighted in red boxes are inverted (figure 4.17). These inversions can also be seen in figure 4.18.

#### **4.6 Summary**

All the Beta-defensin genes resided between the flanking genes CTSB and XPO1. However, their number varied by 82 in *P. muralis* to 34 in *Vivipara*. The majority of the genes show a two exon structure and a few with potentially three exons, with a signal and mature peptide and some exhibited a large pre/pro piece in the first exon. These larger propieces may act as a charge balancer. They seemed to be undergoing a negative/purifying selective direction. Some of the regions exhibited many similar duplicated genes. These followed the 'birth and death' model of gene duplication. Also, the genomic organisation and comparisons showed several different regions within the cluster have inverted. Another event that has caused variation in the separate species. They seem to share very similar orthologous genes starting close to the CTSB gene.

## Chapter 5 - Snakes

### **5. Aims**

Three separate species of snake will be explored and the differences in number of beta-defensin genes present, physical properties and genomic organisation will be discussed. The three different species, from three different families, are The Indian Cobra (*Naja naja*) an Elapidae, The Western Terrestrial Garter Snake (*Thamnophis elegans*) a Colubridae and Prairie Rattlesnake (*Crotalus viridis viridis*) which is from the Viperidae family.

The sequences described in this work were obtained *In silico* through the methods outlined in chapter 2. The beta-defensin genes were annotated based on the six-cysteine motif to establish the complete clusters. These will be described within each species and then compared between species. Also, this work will investigate the drivers of evolution within these clusters of genes describing possible mechanisms for variability. All species in this chapter were used on the basis that the cluster found within the genome assembly was complete.

### **5.1 *Crotalus viridis viridis* – Prairie Rattlesnake**



Photos: Snake - Todd Pierson, <https://reptile-database.reptarium.cz/species?genus=Crotalus&species=viridis>  
Map - <https://www.adaptationenvironmental.com/rattler-tattler-blog/venom-research-in-colorado>

The Prairie Rattlesnake is a venomous pit viper species native to the western United States, southwestern Canada and northern Mexico. The species commonly grows more than 1m in length. The presence of 3 or more internasal scales is a characteristic identifying feature and

coloration is usually different colours of light brown giving good camouflage that allows them to easily blend into their habitat. Darker browns are usually distributed along its dorsal edge. Generally, the Prairie Rattlesnake will occupy areas with abundant prey and tend to prefer dry habitats with some vegetation to allow for cover when hunting. They tend to be ground dwelling snakes but have been known to climb trees. Their preferred prey is small mammals but will occasionally eat amphibians and reptiles. The venom of the Prairie Rattlesnake is a complex mixture of different proteins including hemotoxins that has a tissue destructive ability. The venom also has neurotoxic properties. The snakes are viviparous and can produce up to 25 individuals per reproductive cycle, but numbers may vary due to environmental challenges. They give birth in the late summer with the young being toxic from birth. The species is classed as of Least Concern on the IUCN Red List of threatened species.

#### **4.1.1 Data Mining and Cluster assembly**

The genomic sequences data was accessed through the NCBI genome assembly database. This species was chosen as the assembly was at chromosomal level therefore it was predicted to have the beta-defensin cluster region intact, so when forming the genomic organisation, the complete cluster could be confirmed. The genome GenBank assembly accession number is GCA\_003400415.2 and was submitted to the database on 08/01/2019. As snakes are from the same order of reptiles as lizards, it was hypothesised that the cluster region may reside between CTSB but also XPO1. The cluster was identified and was syntenic with the lizard's clusters in that the beta-defensin genes were discovered to reside between CTSB and XPO1 which was approximately 2.7Mb downstream on chromosome 1 (GenBank sequence CM012306.1).

As described earlier, being syntenic with lizards with regards to the flanking genes, the region identified was searched with concatemers of the beta-defensin amino acid sequences obtained from the lizards in the previous chapter as an initial query using the tBLASTn program on the NCBI server. This region was masked using the RepeatMasker (Smit *et al.* 2006) to remove the repeat sequences from the DNA sequence. This was then translated into a 6-frame output using EMBOSS Sixpack program on The European Bioinformatics Institute (EMBL-EBI) website and this was utilised to highlight potential matches from the tBLASTn concatemer queries.

Using the concatemer produced more matches compared with searches using single amino acid sequences, but does not acquire all the exons and therefore other approaches were employed. Gene finding programs GENSCAN (Burge and Karlin 1997) and FGENESH (Solovyev *et al.* 2006) were employed to search for putative exons that were not initially identified with the tBLASTn approach. Finally, regions of more than 3000bp of the repeats determined in the repeatmasker analysis but not in the vicinity of already resolved exons and downstream from the poly adenylation signal, can then be queried to exclude all potential regions where Beta-defensin exons may reside.

Splice sites were finalised using the online server by the Berkeley Drosophila Genome Project (Reese *et al.* 1997) and amino translations were ascertained from the DNA sequences of predicted exons. Finally, iterative searches were performed using the newly identified beta-defensins against the cluster region.

### **5.1.2 Cluster organisation and Beta-defensin sequences**

In this work a total of 15 predicted beta-defensins were discovered (figure 5.1). These were numbered in order along the chromosome using the flanking gene CTSB as a reference to the start of the cluster as shown in the Komodo dragon (van Hoek *et al.* 2019). The naming of the genes used a prefix to the order number and was an abbreviation of the species name, in this case CVBD. The cluster was located on chromosome 1 (Genbank sequence CM012306.1) of the genome assembly (GenBank assembly accession: GCA\_003400415.2) between positions 289557092-292248510 for which the reverse complement was used to start the cluster at the CTSB gene. Relative positions and genomic organisation are shown in figure 5.2 as well as other information in appendix 2.1.

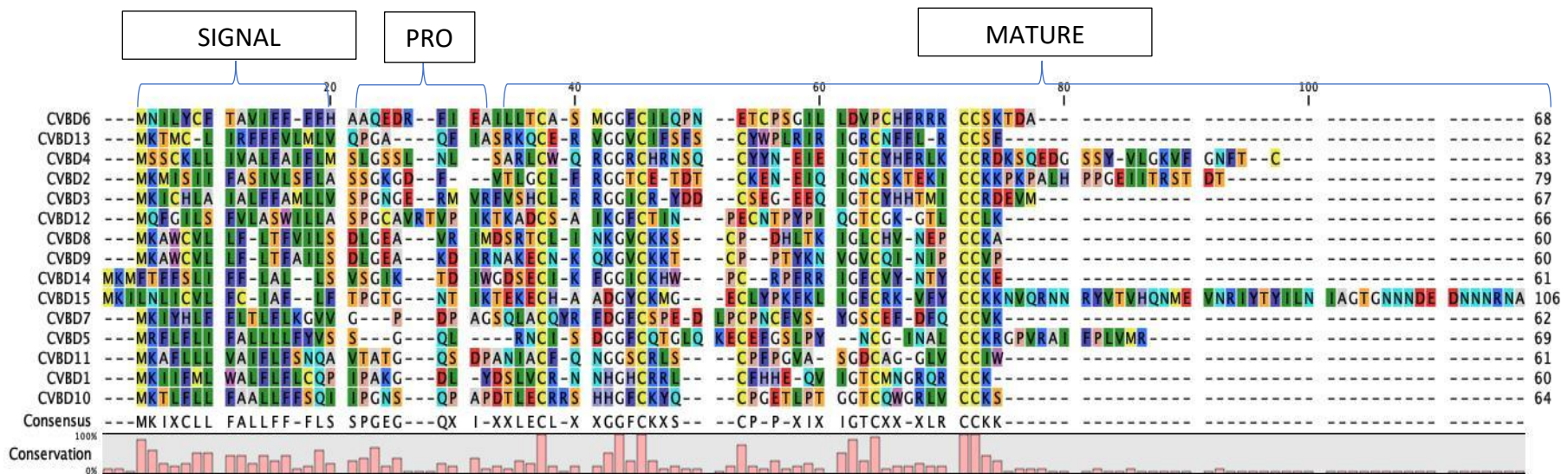
The last beta-defensin in this region was located around 313kb long from the last codon of CTSB. This made the length of sequence between the last beta-defensin and XPO1 approximately 2Mb. There appeared to be an 'empty' region within this part of the genomic sequence which did not contain any beta-defensins. BLAST searches were performed by using the DNA sequence of this region in the BLASTx program using reptiles and birds as reference organisms to see if any other possible genes were present within this region. No significant results were obtained, and more investigation is needed to fully assess why this region is present.

The beta-defensins discovered in this analysis show the conserved structure consisting of two exons. The first exon encodes a signal peptide followed by the second exon, encoding the mature peptide consisting of the typical defensin motif with common 6 cysteine domain with a glycine located in the N-terminal region, two positions upstream from the second cysteine, while another glycine is placed two or three positions upstream from the fourth cysteine. The rest of the amino acids in the domain are less conserved.

### **5.1.3 Physical Properties**

The physical properties (Appendix 2.2) of the beta-defensins identified also show a degree of diversity. Beta-defensins are usually described as cationic but unexpectedly the mature peptides of CVBD2 had no charge and CVBD3, 7 and 11 holding a negative charge. This could suggest a yet unknown function of these peptides. The molecular weight of the beta-defensins vary from Mr of 4312 in CVBD8 to Mr of 9818 of CVBD15. Signal peptides are a short amino acid sequence in the n-terminus of many newly synthesised proteins, and these serve as a target to allow proteins to be processed into or across the cell membrane. All the beta-defensins found in this cluster have a signal peptide which is typical of defensins (appendix 2.3). Confirmed using SignalIP – 5.0 server (Almagro Armenteros *et al* 2019)





**Figure 5.1 Multiple sequence alignment of *Crotalus viridis viridis* beta-defensins.**

Produced in Clustal X. The beta-defensin genes in this cluster show the typical conserved signal peptide at the start of the gene and in the mature peptide showing the 6 conserved cysteine residues along with glycine residues with a small pro peptide situated in between. Conservation percentages are shown underneath the alignment. Signal, Pro-peptide and Mature regions are shown by the parentheses



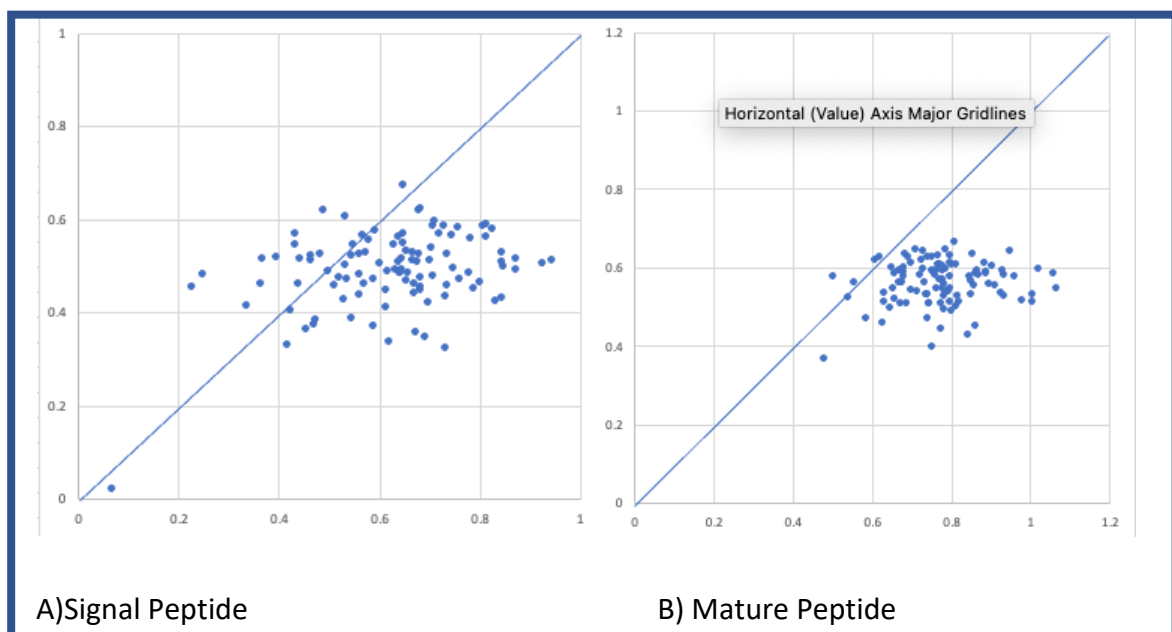
**Figure 5.2 Genomic organisation of the Beta-defensin cluster of *Crotalus viridis viridis*.**

A total of 15 genes were discovered. Each vertical line represents 100kb along the chromosome. The blue blocks represent the full genes, and the arrows dictate their orientation. The two diagonal lines show that XPO1 gene is upstream from this and not the full length of the cluster. Each vertical line represents 100kb. The yellow illustrates the flanking gene CTSB at the start of the cluster. The size of the squares (genes) or the distance between them are only representative and are not proportional to their actual size in the genome.



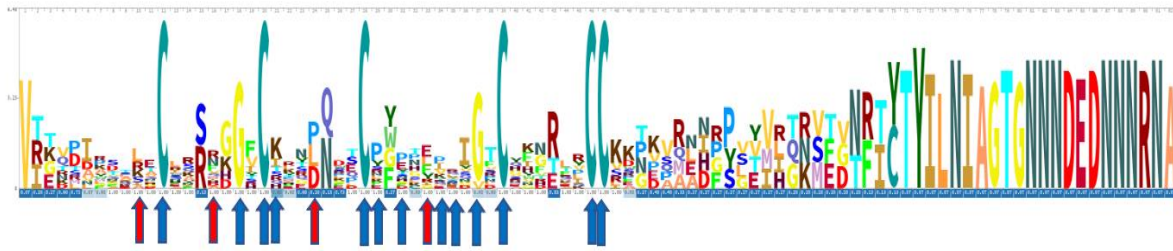
### 5.1.4 Selection Analyses

Pairwise comparisons and site wise selection analyses were performed on the beta-defensin genes in the cluster (figure 5.3) as laid out in section 3.1.4. Given that beta-defensin clusters arise from gene duplication and paralogous to each other, pairwise comparisons of each gene against the next were performed. The trends that are observable in the data suggest that both the signal peptide and the mature peptide are undergoing slight purifying selection, however, the signal peptide shows more points of nonsynonymous substitutions than the mature peptide. The proportions were more purifying for the mature peptide. This observation could be down to the conserved cysteine and glycine residues within the mature peptide.



**Figure 5.3 Proportion of synonymous and nonsynonymous substitutions in *C. v. viridis*.**

Ratios within the signal peptide (A) and the mature peptide (B) with nonsynonymous ( $d_N$ ) on the y axis and synonymous ( $d_S$ ) on the x axis. The diagonal lines represent  $d_N = d_S$  and is given for estimating selection pressures; dots above and below this line represent positive and purifying selection, respectively.



**Figure 5.4 Amino acid sequence logo of mature peptides of *C. v. viridis*.**

Sites which are undergoing positive selection (red arrow) and purifying selection (blue arrow) tested by FEL, MEME and FUBAR in HYPHY. Logo produced on Skylign.org.

Selection analysis of the individual amino acid sites within the peptide was performed using HyPhy (Pond *et al* 2005; Pond *et al* 2005a) (Figure 5.4). It is observed that the conserved regions show purifying selection especially within the cysteine motif but there are sites between these conserved regions showing positive selection. A total of 4 sites were identified as being under positive selection. Where the cysteine bonds in the peptide are formed, the 'bends' between these are where positive selection is occurring. This could play a role in diversity and changing the shape of the tertiary structure of the peptide and giving an array of shapes within the genes of the cluster, providing a greater degree of protection against pathogens by providing an arsenal of different shaped peptides.

### **5.1.5 Repeat Sequence landscape**

Repeat masker was performed using query species database set to tetrapod. The *C. v. viridis* defensin cluster region had over all 17.29% bases masked with the predominant repeat elements being retroelements at 77.8% of bases masked. LINES were around 78.5% of the retroelements and CR1 LINE being the most abundant at 62.2% of the LINES present. LTR elements accounted for about 15.9% of the retroelements. Around 8.8% of the repeat sequences were DNA transposons with Hobo-Activator being the most abundant (table 5.1).

```

=====
file name: Crotalus_Viridis_CTSB-XP01.fa
sequences: 1
total length: 2757929 bp (2648515 bp excl N/X-runs)
GC level: 37.98 %
bases masked: 476803 bp ( 17.29 %)
=====

```

	number of elements*	length occupied	percentage of sequence
Retroelements	1123	376050 bp	13.64 %
SINEs:	193	20707 bp	0.75 %
Penelope	60	8859 bp	0.32 %
LINEs:	813	295560 bp	10.72 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	487	183916 bp	6.67 %
R1/LOA/Jockey	1	100 bp	0.00 %
R2/R4/NeSL	55	22293 bp	0.81 %
RTE/Bov-B	108	44797 bp	1.62 %
L1/CIN4	100	35461 bp	1.29 %
LTR elements:	117	59783 bp	2.17 %
BEL/Pao	0	0 bp	0.00 %
Ty1/Copia	2	1196 bp	0.04 %
Gypsy/DIRS1	59	54032 bp	1.96 %
Retroviral	56	4555 bp	0.17 %
DNA transposons	440	42333 bp	1.53 %
hobo-Activator	326	24511 bp	0.89 %
Tc1-IS630-Pogo	68	13536 bp	0.49 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	16	1677 bp	0.06 %
Other (Mirage, P-element, <u>Transib</u> )	0	0 bp	0.00 %
Rolling-circles	1	64 bp	0.00 %
Unclassified:	15	2104 bp	0.08 %
Total interspersed repeats:		420487 bp	15.25 %
Small RNA:	7	508 bp	0.02 %
Satellites:	30	7011 bp	0.25 %
Simple repeats:	936	39923 bp	1.45 %
Low complexity:	166	8992 bp	0.33 %

\* most repeats fragmented by insertions or deletions  
have been counted as one element

The query species was assumed to be tetrapods  
RepeatMasker version 4.1.2-p1 , default mode

run with rmblastn version 2.2.27+  
FamDB: CONS-Dfam withRBRM 3.3

**Table 5.1 Repeat masker summary for *C. v. viridis*.**

*Displaying the different repeat sequences within the C. v. viridis cluster region. The tetrapod database was used as a reference for the repeat sequence matches in RepeatMasker program.*

## 5.2 *Naja naja* – The Indian Cobra



Photos: Snake - <https://www.techexplorist.com/scientists-decoded-genome-indian-cobra/28903/>  
Map - <https://a-z-animals.com/animals/indian-cobra/>

The Indian Cobra is a venomous elapid snake native to India and the surrounding countries including Pakistan, Bangladesh, Sri Lanka, Nepal, and Bhutan and has been made famous as its often seen with snake charmers in this region. The Indian Cobra is easily identified by its impressive hood which expands when threatened. Many possess a hood marking which is located behind its head and resembles that of spectacles with two connected circular patterns. They grow up to 1.5 meters in length, but specimens have been found up to 2.2 meters long. There is much variation in their colouration depending on where they are found and can be grey, yellow, tan, brown or black. The Indian Cobra's habitat consists of a wide range of environments and can be often found near water. They inhabit dense forest, plains and agricultural lands including paddy fields, rocky terrain, and wetlands. It is not found in high altitudes above 2000m and extreme desert regions. Indian Cobras are oviparous and lay their eggs between the months of April and July. They can lay up to 30 eggs which hatch 48-69 days later. The young have fully functioning venom glands. The Indian Cobra venom is a powerful post-synaptic neurotoxin and a cardiotoxin. It acts on the synapses of nerves causing paralysis and cardiac arrest. It is protected under the Indian Wildlife Protection Act 1972.

### **5.2.1 Data mining and cluster assembly**

The genomic sequence data was accessed through the NCBI genome assembly database. This species was chosen as the assembly was at chromosomal level and there would be confidence

that it would contain the complete beta-defensin cluster region. The GenBank assembly accession number is GCA\_009733165.1 and was submitted to the database on 11/12/2019. As with the previous analysis, the region identified was searched with concatemers of the Beta-defensin amino acids sequences obtained from the lizards. As XPO1 was identified in the *C. v. viridis* genome as the flanking gene of the cluster region this was also used as a query to establish the region containing potential beta-defensins.

The previous amino acid concatemer sequences along with the beta-defensins discovered in the *C. v. viridis* defensin cluster were used as a query search within this region using the tBLASTn program on the BLAST server. This region was masked using the RepeatMasker (Smit *et.al.* 2006) to remove the repeat sequences from the DNA sequence. This was then translated into a 6-frame output using EMBOSS Sixpack program on The European Bioinformatics Institute (EMBL-EBI) website and this was utilised to highlight potential matches.

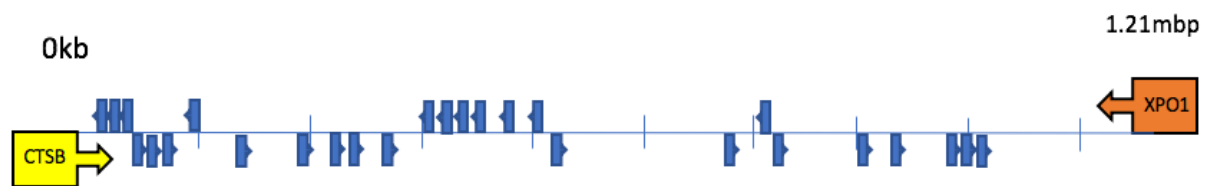
This process, once again, produced more matches when compared with searches using single amino acid sequences, but it does not acquire all the exons and therefore other approaches were employed. Firstly, the DNA coding sequences discovered in the *C. v. viridis* genome were used as a query using the BLASTn program. This gave matches of close orthologues of the beta-defensins found. Secondly, gene finding programs GENSCAN (Burge and Karlin 1997) and FGENESH (Solovyev *et al.* 2006) were employed to search for putative exons that weren't initially found with the BLAST approach. Finally, regions of more than 3000bp of the repeats determined in the RepeatMasker analysis but not in the vicinity of already resolved exons and downstream from the poly adenylation signal can then be searched to exclude all potential regions where beta-defensin exons may reside.

Splice site prediction was finalised using the online server by the Berkeley Drosophila Genome Project (Reese *et al.* 1997) and amino translations were ascertained from the DNA sequences of potential exons. Finally, iterative searches were performed using the newly identified beta-defensins against the cluster region.

### 5.2.2 Cluster organisation and Beta-defensin sequences

The cluster was found between positions 32523301-33738222 on chromosome 1 (GenBank sequence CM019148.1) of the genome assembly (GenBank assembly accession: GCA\_009733165.1). This analysis on this cluster revealed a total of 27 beta-defensins (appendix 2.4) located between CTSB and XPO1. The length of the cluster was approximately 1.21mbps (figure 5.5). The naming of the genes used a prefix to the order number and was an abbreviation of the species name, in this case NNBD.

Again, the beta-defensins discovered in this analysis show the conserved structure consisting of two exons. The first exon encodes a signal peptide followed by the second exon, encoding the mature peptide consisting of the typical defensin motif with common 6 cysteine domain with a glycine located in the N-terminal region, two positions upstream from the second cysteine, while another glycine is placed two or three positions upstream from the fourth cysteine. The rest of the amino acids in the domain are less conserved.



**Figure 5.5 Genomic organisation of the *Naja naja* beta-defensin cluster.**

A total of 27 genes were found between CTSB and XPO1. Arrows indicate which direction they are found in relation to CTSB. Each vertical line represents 100kb. The yellow and orange boxes illustrate the flanking genes CTSB and XPO1 respectively. The size of the squares (genes) or the distance between them are only representative and are not proportional to their actual size in the genome.







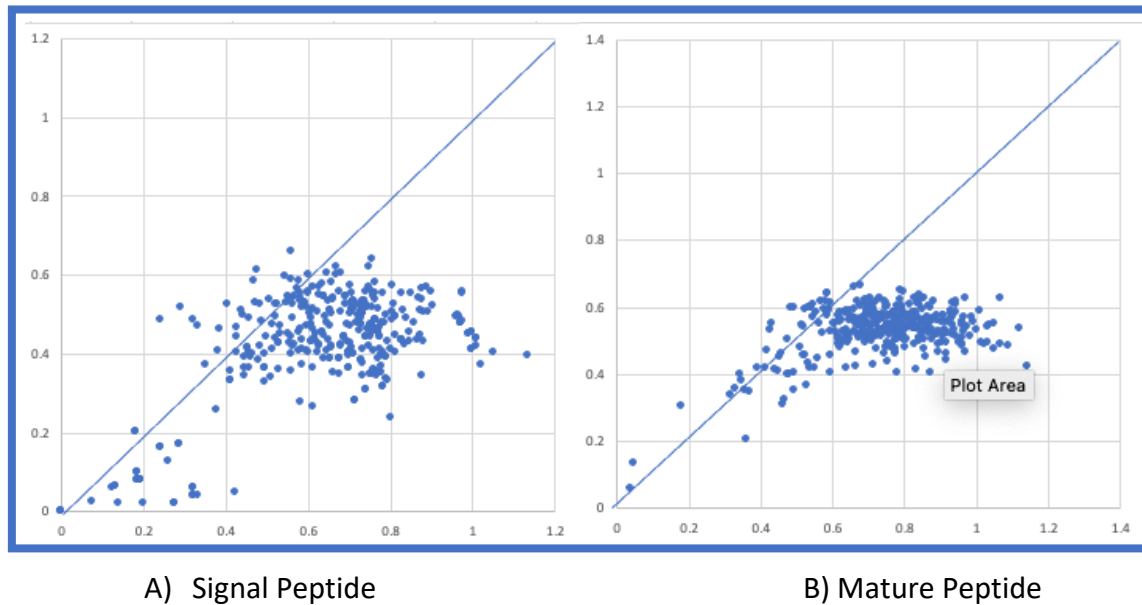
**Figure 5.7 Phylogenetic tree of the DNA coding sequences of exons 1 and 2 of *Naja naja*.** Produced in the IQ-tree server (Trifinopoulos et al 2016) using ultrafast analysis of 1000 bootstrap alignments. The red highlighted genes show recent duplications with high degree of conservation.

### 5.2.3 Selection Analyses

Exploring the phylogeny in figure 5.7 NNBD20-27 (highlighted in red) suggests that their duplications happened more recently compared to the other beta-defensins in the cluster. This is also confirmed by the similarity of the paralogue sequences that are shown in the multiple sequence alignment. These defensins also fit the ‘birth and death’ model of evolution of duplicated genes as described by Nei and Hughes (1992). This model describes two main features 1) an interspecific gene clustering pattern and 2) the presence of pseudogenes (Eirín-



López *et al.* 2012). In addition, when comparing to the *C. v. viridis* beta-defensins they do not show conservation of synteny, so these could have arisen later in the evolution of the Indian Cobra. More on this will be explored later in the chapter.



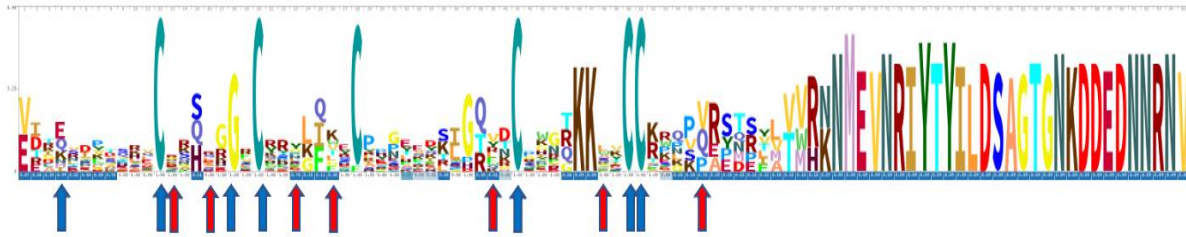
**Figure 5.8 Proportion of synonymous and nonsynonymous substitutions in *Naja naja*.**

Ratios in signal peptide (A) and the mature peptide (B) are shown as nonsynonymous ( $d_N$ ) on the y axis and synonymous ( $d_S$ ) on the x axis. The diagonal lines represent  $d_N = d_S$  and is given for estimating selection pressures; dots above and below this line represent positive and purifying selection, respectively.

As with the *C. v. viridis* beta-defensins, evolutionary analyses were also conducted on the beta-defensin genes in the cluster. Given that beta-defensin clusters arise from gene duplication and therefore paralogous to each other pairwise comparisons of each gene against the next looking into the proportion of non-synonymous ( $d_N$ ) to synonymous ( $d_S$ ) substitutions. This was done separately for the signal peptide and the mature peptide (figure 5.8).

The trends observed in the data suggest that the signal peptide has a greater proportion of synonymous substitutions showing a degree of purifying selection. The mature peptide also

displays a large distribution of synonymous substitutions implying purifying selection; however, this may be due to the recent duplications (CV20-27) having such similar sequences. Selection analysis was performed using HyPhy (Pond *et al* 2005) via the datamonkey online server (Pond *et al* 2005a). The mature peptide was analysed to see which individual amino acid sites were undergoing selection (Figure 5.9)



**Figure 5.9 Amino acid sequence logo of mature peptides in *Naja naja*.**

Sites which are undergoing positive selection (red arrow) and purifying selection (blue arrow) as tested by FEL, MEME and FUBAR using in HYPHY. Logo produced on Skylign.org

Analysis shows that the conserved regions of the Beta-defensin cysteine motif follows purifying selection but the regions between these are showing positive selection. A total of 7 sites were identified as undergoing positive selection. These were situated between the conserved sites and may play a role in diversifying the tertiary structure. The logo also illustrates the purifying sites notably the conserved defensin cysteines.

#### 5.2.4 Physical Properties

The physical properties of each of the mature Beta-defensin peptides shows a good array of different and varying features (appendix 2.5). As with the *C. v. viridis* beta-defensins there are three beta-defensins that are anionic – NNBD3, 7, 8, 12 and 14 along with one with no charge, NNBD10. NNBD14 has a charge of -5 and this may be an interesting candidate for further investigation. NNBD9 is also quite cationic having a charge of 8. This would also be of interest for further investigation as it may have a yet, unknown function. The molecular weights of the peptide vary 9785-3911. All properties were achieved by using the protparam program on the ExPASy Server (Gasteiger, E. *et al.* 2005).

All the beta-defensins found in this cluster have a signal peptide which is typical of defensins. Performed using SignalIP – 5.0 server (Almagro Armenteros *et al* 2019) (appendix 2.6)

### 5.2.5 Repeat Sequence landscape

Repeat masker was performed using query species database set to tetrapod. The *N. naja* defensin cluster region had over all 32.92% bases masked with the predominant repeat elements being retroelements at 90.4% of bases masked. LINES were around 76.8% of the retroelements and CR1 LINE being the most abundant at 75.8% of the LINES present. LTR elements accounted for about 20.9% of the retroelements. Around 3% of the repeat sequences were DNA transposons with Hobo-Activator being the most abundant (table 5.2).

```

=====
file name: Naja_naja_DNA_cluster.fa
sequences: 1
total length: 1214922 bp (1207922 bp excl N/X-runs)
GC level: 41.85 %
bases masked: 399928 bp ( 32.92 %)
=====

```

	number of elements*	length occupied	percentage of sequence
Retroelements	734	361685 bp	29.77 %
SINEs:	99	7981 bp	0.66 %
Penelope	19	2785 bp	0.23 %
LINES:	553	277911 bp	22.87 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	411	210773 bp	17.35 %
R1/LOA/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	5	1928 bp	0.16 %
RTE/Bov-B	46	17775 bp	1.46 %
L1/CIN4	70	44489 bp	3.66 %
LTR elements:	82	75793 bp	6.24 %
BEL/Pao	0	0 bp	0.00 %
Ty1/Copia	3	750 bp	0.06 %
Gypsy/DIRS1	45	68538 bp	5.64 %
Retroviral	34	6505 bp	0.54 %
DNA transposons	110	12333 bp	1.02 %
hobo-Activator	76	5336 bp	0.44 %
Tc1-IS630-Pogo	24	6385 bp	0.53 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	1	72 bp	0.01 %
Tourist/Harbinger	1	20 bp	0.00 %
Other (Mirage, P-element, <u>Transib</u> )	0	0 bp	0.00 %
Rolling-circles	1	80 bp	0.01 %
Unclassified:	20	1355 bp	0.11 %
Total interspersed repeats:		375373 bp	30.90 %
Small RNA:	4	312 bp	0.03 %
Satellites:	6	803 bp	0.07 %
Simple repeats:	443	19579 bp	1.61 %
Low complexity:	74	4021 bp	0.33 %

```

=====
* most repeats fragmented by insertions or deletions
  have been counted as one element

The query species was assumed to be tetrapods
RepeatMasker version 4.1.2-p1 , default mode

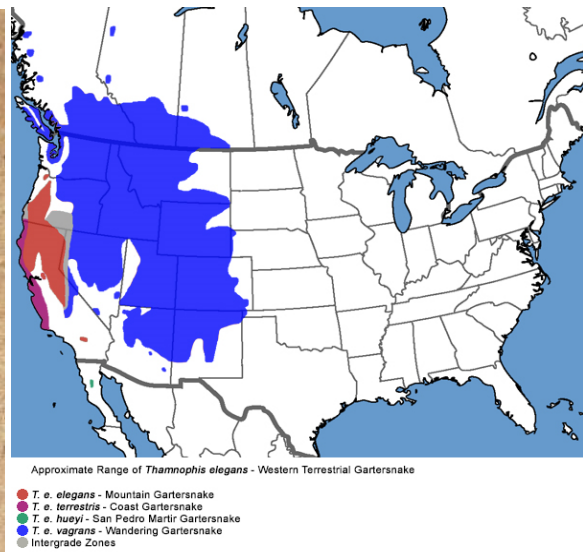
run with rmblastn version 2.2.27+
FamDB: CONS-Dfam_withRBRM_3.3
=====

```

**Table 5.2 Repeat masker summary for *Naja naja*.**

Displaying the different repeat sequences within the *N. naja* cluster region. The tetrapod database was used as a reference for the repeat sequence matches in RepeatMasker program.

### 5.3 *Thamnophis Elegans* - The Western Terrestrial Garter Snake



Photos: Snake - J. N. Stuart, <https://www.inaturalist.org/photos/3624>

Map - <http://www.californiaherps.com/snakes/pages/t.e.vagrans.html>

The Western Terrestrial Garter Snake is a north American species of colubrid snake with six subspecies identified. They are found in some parts of Canada as well as western Nebraska and Oklahoma in the US. They are medium-sized snakes, usually 46–104 cm (18–41 in). The colouring of the Western Terrestrial Garter Snake varies considerably but usually exhibit a yellow, light orange or white dorsal strip with two side stripes of the same colour. They inhabit a wide variety of different habitats which include woodlands, coniferous forests, and grasslands and can be found at sea level and high altitudes up to 13000 feet. It is primarily ground dwelling and is also semi aquatic. They possess mildly venomous saliva and are believed to be the only garter snake species that constrict prey, although, this constriction is inefficient when compared to other constrictors (de Queiroz & Groen 2001). Their diet depends on what is available in the environment they reside with two main variants: coastal or inland. Coastal populations rely on a diet of slugs, amphibians, small mammals, and lizards. Whereas the inland variant is a semi-aquatic diet and consists of amphibians, leeches, and fish. The Western Terrestrial Garter Snake does not lay eggs but instead is ovoviviparous, which is typical of natricine snakes and will have up to 12 young, which are born between August and September.

### 5.3.1 Data mining and cluster assembly

The genomic sequences data was accessed through the NCBI genome assembly database. This species was again, chosen as the assembly was at chromosomal level therefore it would likely have the beta-defensin cluster region intact allowing a full in-depth search of the beta-defensins present in the area. The GenBank assembly accession number is GCA\_009769535.1 and was submitted to the database on 23/12/2019. Using CTSB as the reference for the start of the analysis, the chicken CTSB amino acid sequence was used as a query to search against the genome using the tBLASTn program. As XPO1 was identified in the *C. v. viridis* and the *N. naja* genomes as the flanking gene of the cluster region it was likely that this would also apply for the *T. elegans* genome, so this was used to establish the region to be further analysed for the beta-defensins.

Using the amino acid sequences obtained whilst determining the beta-defensins discovered in the *C. v. viridis* and *N. naja* defensin cluster, concatemers were produced. These concatemers were then used as a query search within this region using the tBLASTn program on the BLAST server. This region was masked using the RepeatMasker (Smit *et al.* 2006) to remove repeat sequences from the DNA sequence. This was then translated into a 6-frame output using EMBOSS Sixpack program on The European Bioinformatics Institute (EMBL-EBI) website and this was utilised to highlight potential matches and for further analysis.

Using the concatemer approach resulted in more matches when compared to searches using single amino acid sequences, but it does not acquire all the exons and therefore additional approaches were employed. Despite this there was a greater degree of success when using a concatemer of closely related species, such as snakes.

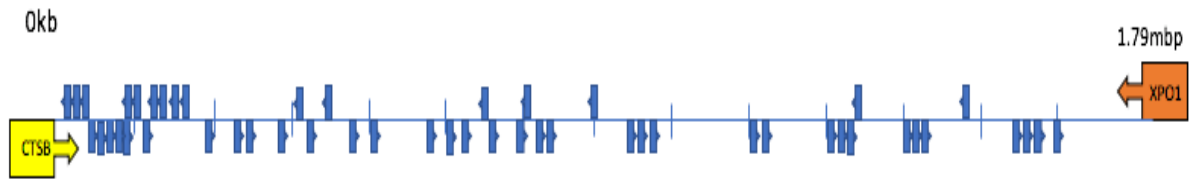
As well as using the above method, the DNA coding sequences were also used as a query using the BLASTn program. This gave matches of close orthologues of the beta-defensins found in the *C. v. viridis* and *N. naja* cluster regions. Additionally, gene finding programs GENSCAN (Burge and Karlin 1997) and FGENESH (Solovyev *et al.* 2006) were employed to search for putative exons that were not initially found with the BLAST approach. Finally, regions of more than 3000bp of the repeats determined in the RepeatMasker analysis but not in the vicinity of already resolved exons and downstream from the poly adenylation signal can then be searched to exclude all potential regions where Beta-defensin exons may reside.

Splice site prediction was finalised using the online server by the Berkeley Drosophila Genome Project (Reese *et al.* 1997) and amino translations were ascertained from the DNA sequences of potential exons. Finally, iterative searches were performed using the newly identified beta-defensins against the cluster region.

### **5.3.2 Cluster organisation and Beta-defensin sequences**

The cluster can be found between positions 108661433-110450888 on chromosome 4 (GenBank sequence CM020099.1) of the genome assembly. This sequence was reversed so that the order was in line with CTSB to start the cluster. This analysis revealed a total of 51 beta-defensins (appendix 2.7) located between the CTSB and XPO1 gene. The length of the cluster is approximately 1.79Mbps. Gene nomenclature is the same as the other genes with the abbreviation being the first two initials followed by the number or order starting at CTSB and in this case TEBD.

The beta-defensins that were discovered in this analysis show the classical two exon gene structure. The first being a signal peptide followed by the second exon being the mature peptide that consists of the typical defensin motif with common 6 cysteine domain and a glycine located in the N-terminal region, two positions upstream from the second cysteine, while another glycine is placed two or three positions upstream from the fourth cysteine. The rest of the amino acids in the domain are poorly conserved. Figure 5.9 shows the genomic organisation and figure 5.11 is the multiple sequence alignment showing the conservation typical in the sequence of beta-defensins.

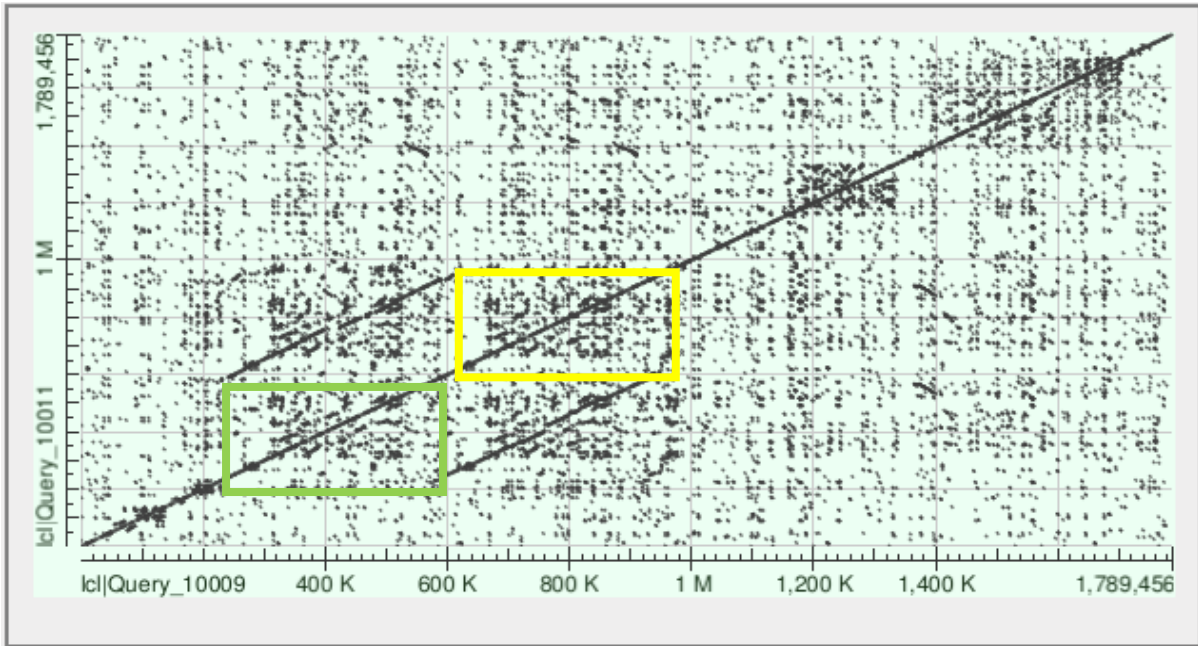


**Figure 5.9 Genomic organisation of the *Thamnophis elegans* beta-defensin cluster.**

A total of 51 genes were found between CTSB and XPO1. Arrows indicate which direction they are found in relation to CTSB. Each vertical line represents 100kb. The yellow and orange boxes illustrate the flanking genes CTSB and XPO1 respectively. The size of the squares (genes) or the distance between them are only representative and are not proportional to their actual size in the genome. Double diagonal line showing the end of one scaffold and start of the next.

The dot plot illustrates a large cluster duplication. A region from around 250,000 bp to 580,000 bp has duplicated to the region from 580,000 bp to 950,000 bp or vice versa. This can also be seen in sequences where homology is shown in figure 5.10 and table 5.3 where the sequences highlighted in green have duplicated with the sequences highlighted in yellow. There are also regions of high duplication within the cluster that also can be seen in the phylogenetic tree and the sequences of the defensins themselves. One thing to note is the region towards the end of the cluster. This is also shown as blue boxes in the synteny diagram (figure 5.16).

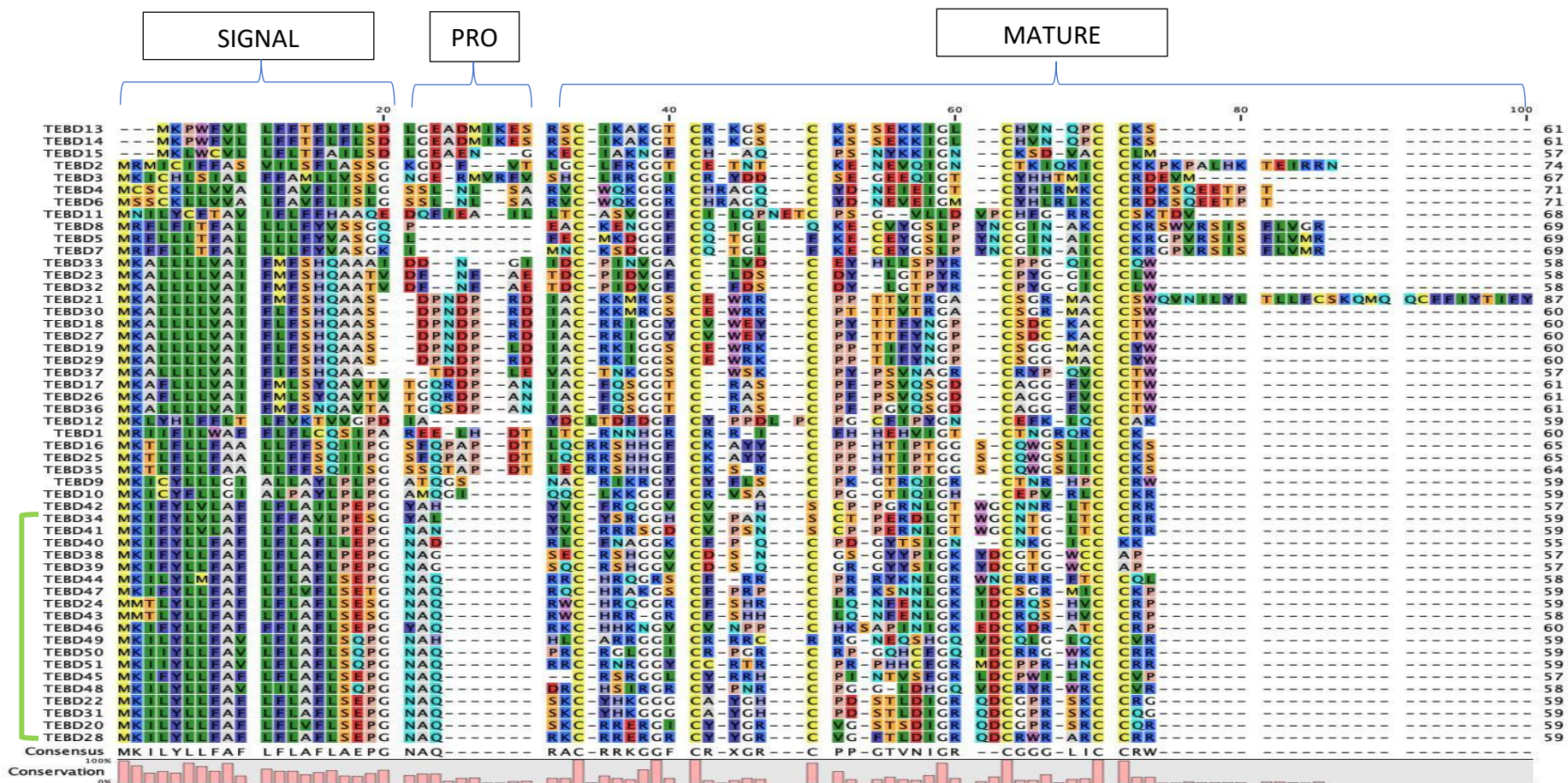




**Figure 5.10** Dot plot of the cluster region of *T. elegans*

Each dot and line is a point of homology with other parts of the cluster. The green and yellow boxes show regions of a cluster duplication.





**Figure 5.11 Multiple sequence alignment of *Thamnophis Elegans* beta-defensins.**

Produced using Clustal X. It is clear to see that the beta-defensin genes in this cluster show the typical conserved signal peptide at the start of the gene and in the mature peptide showing the 6 conserved cysteine residues along with glycine residues with a small pro peptide situated in between. Conservation percentages are shown underneath the alignment. Signal, Pro-peptide and Mature regions are also shown by the parentheses. The green bracket highlights highly similar paralogs that are shown as outlier group in figure 5.12.

GENE	SIGNAL PEPTIDE	MATURE PEPTIDE
TEBD16	MKTLFLLFAALLFFSQIIPG	SFQPAPDTLQCRRSHHGFCAYYCPHTIPTGGSCQWWSLICCKS
TEBD17	MKAFLLLVAFMFLSYQAVTVTG	QRDPANIACFQSGGTCRASCFFPSVQSGDCAGGFVCTW
TEBD18	MKALLLVAFIFLFSHQAAAS	DPNDPRDIACRRIGGYCVWEYCPYTYFYNGPCSDCKACCTW
TEBD19	MKALLLVAFIFLFSHQAAAS	DPNDPLDIACRKIGGSCEWRKCPPTIFYNGPCSGGMACCYW
TEBD20	MKILYLLFAFLFLVFLSEPGNA	QSKCRRERGCYGRVGVSTSDIGRQDCGPRSRCCQR
TEBD21	MKALLLVAFIFMFSHQAAAS	DPNDPRDIACKMRGSCSEWRRCPTTVTRGACSGRMACCSWQVNIYLTLFLCSKQMQCCFFIYTFY
TEBD22	MKILYLLFAFLFLAFSEPGNA	QSKCYHKGGCAYGHCPDSTLDIGRQDCGPRSKCCRG
TEBD23	MKALLLVAFIFMFSHQAAAT	VDFNFAETDCPIDVGFCLDSCDYLGTPYRCPYGGICCLW
TEBD24	MMTLYLLFAFLFLAFSESGNA	QRWCHRQGGRCFSHRCLQNFENLGKIDCRQSHVCCRP
TEBD25	MKTLFLLFAALLFFSQIIPG	SFQPAPDTLQCRRSHHGFCAYYCPHTIPTGGSCQWWSLICCKS
TEBD26	MKAFLLLVAFMFLSYQAVTVTG	QRDPANIACFQSGGTCRASCFFPSVQSGDCAGGFVCTW
TEBD27	MKALLLVAFIFLFSHQAAAS	DPNDPRDIACRRIGGYCVWEYCPYTYFYNGPCSDCKACCTW
TEBD28	MKILYLLFAFLFLAFSEPGNA	QRKCRREGRGCYGRVGVFTLDIGRQDCRWRARCCRR
TEBD29	MKALLLVAFIFLFSHQAAAS	DPNDPRDIACRKIGGSCEWRKCPPTIFYNGPCSGGMACCYW
TEBD30	MKALLLVAFIFLFSHQAAAS	DPNDPRDIACKMRGSCSEWRRCPTTVTRGACSGRMACCSW
TEBD31	MKILYLLFAFLFLAFSEPGNA	QSKCYHKGGCAYGHCPDSTLDIGRQDCGPRSKCCQG
TEBD32	MKALLLVAFIFMFSHQAAAT	VDFNFAETDCPIDVGFCLDSCDYLGTPYRCPYGGICCLW
TEBD33	MKALLLVAFIFMFSHQAAA	IDDNGIIDCPINVGACLDCEYHLLSPYRCPGQICQW
TEBD34	MKIFYLVLAFLFFAVLPESGYA	LYLCYSRGGHCVPANSCTPERDLGTWGCNTGLTCCRR

**Table 5.3 Signal peptide cleavage sites in the Beta-defensin cluster for *T. elegans*.**

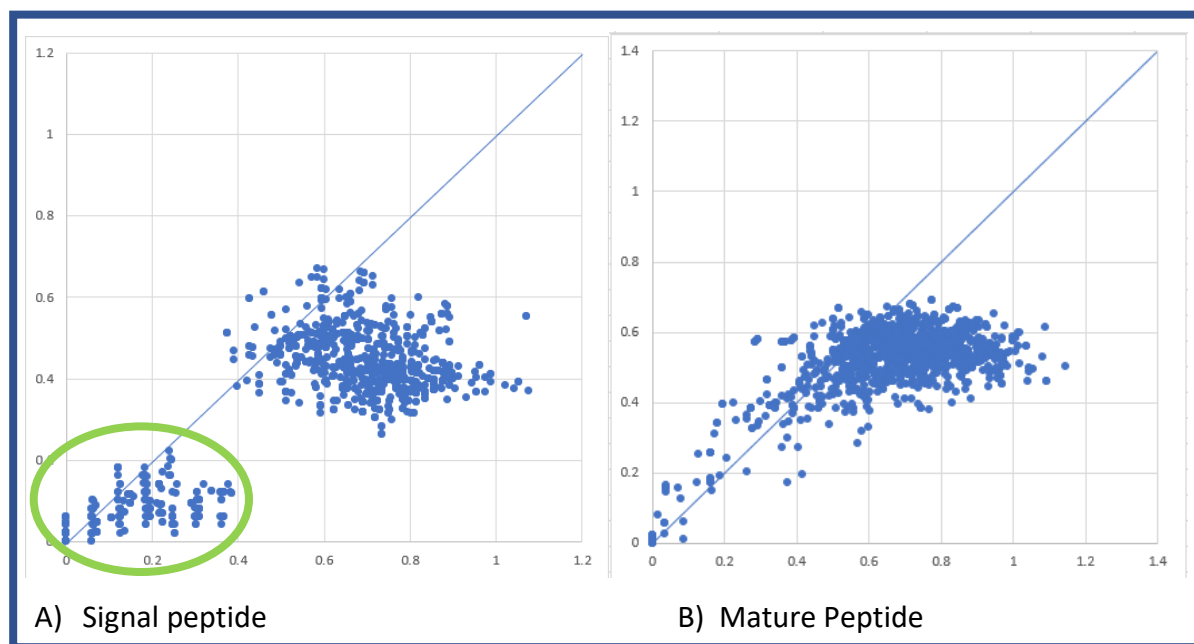
Showing signal sequence and mature peptide sequence. Yellow and green highlighted sequences indicate the cluster duplication shown in figure 5.10.

### 5.3.3 Physical Properties

The physical properties (appendix 2.8) of each of the mature Beta-defensin peptides show a good collection of different and varying features. As with the *C. v. viridis* beta-defensins there are beta-defensins that are anionic – TEBD3, 12, 18, 23, 27, 32, 33, 36 and 38 along with 3 that has no charge – TEBD11, 17 and 26. TEBD44 is also very cationic having a charge of 13. The molecular weights of the peptide vary between 7954-3542 Mr. All properties were achieved using the protparam program on the ExPASy Server (Gasteiger, E. *et al.* 2005). All of the beta-defensins found in this cluster have a signal peptide which is typical of defensins. (appendix 2.9) Performed using SignalIP – 5.0 server (Almagro Armenteros *et al* 2019).

### 5.3.4 Selection analyses

Selection analyses were also conducted on the beta-defensin genes in the *T. elegans* cluster (figure 5.12). In the signal peptide there is an even distribution of synonymous and nonsynonymous substitutions which may be an indicator that they are undergoing slight diversification. The mature peptides show a greater degree of synonymous substitutions showing purifying selection, again indicating that this is due to the number of recent duplications showing a high degree of homology and therefore are seen to be undergoing purifying selection. The highlighted green circle shows that a group of highly similar duplicated genes are present and shown in the multiple sequence alignment.



**Figure 5.12 Ratio of synonymous and nonsynonymous substitutions in *T. elegans*.**

Ratios within the signal peptide (A) and the mature peptide (B) are shown nonsynonymous ( $d_N$ ) on the y axis and synonymous ( $d_S$ ) on the x axis. The diagonal lines represent  $d_N = d_S$  and is given for estimating selection pressures; dots above and below this line represent positive and purifying selection, respectively. The green circle is highlighting the paralogues shown by a green parenthesis on the multiple sequence alignment (figure 5.11). The peptides show a high degree on similarity.





```

=====
file name: Thamnophis_elegans_DNA_cluster.fa
sequences: 1
total length: 1789456 bp (1789456 bp excl N/X-runs)
GC level: 40.29 %
bases masked: 480297 bp ( 26.84 %)
=====

```

	number of elements*	length occupied	percentage of sequence
Retroelements	948	404191 bp	22.59 %
SINEs:	170	15881 bp	0.89 %
Penelope	46	6235 bp	0.35 %
LINEs:	681	343581 bp	19.20 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	390	180279 bp	10.07 %
R1/LOA/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	48	67343 bp	3.76 %
RTE/Bov-B	64	20674 bp	1.16 %
L1/CIN4	133	69050 bp	3.86 %
LTR elements:	97	44729 bp	2.50 %
BEL/Pao	1	74 bp	0.00 %
Ty1/Copia	4	649 bp	0.04 %
Gypsy/DIRS1	34	33087 bp	1.85 %
Retroviral	58	10919 bp	0.61 %
DNA transposons	170	26083 bp	1.46 %
hobo-Activator	105	9138 bp	0.51 %
Tc1-IS630-Pogo	42	10805 bp	0.60 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	1	63 bp	0.00 %
Tourist/Harbinger	6	224 bp	0.01 %
Other (Mirage, P-element, <u>Transib</u> )	0	0 bp	0.00 %
Rolling-circles	8	641 bp	0.04 %
Unclassified:	10	543 bp	0.03 %
Total interspersed repeats:		430817 bp	24.08 %
Small RNA:	2	141 bp	0.01 %
Satellites:	25	3489 bp	0.19 %
Simple repeats:	762	35635 bp	1.99 %
Low complexity:	140	9670 bp	0.54 %

```

=====
* most repeats fragmented by insertions or deletions
  have been counted as one element

The query species was assumed to be tetrapods
RepeatMasker version 4.1.2-p1 , default mode

run with rmblastn version 2.2.27+
FamDB: CONS-Dfam withRBRM_3.3
=====

```

**Table 5.4 Repeat masker summary in *T. elegans*.**

Displaying the different repeat sequences within the *T. elegans* cluster region. The tetrapod database was used as a reference for the repeat sequence matches in RepeatMasker program.

#### 5.4 Conservation of synteny analysis between species

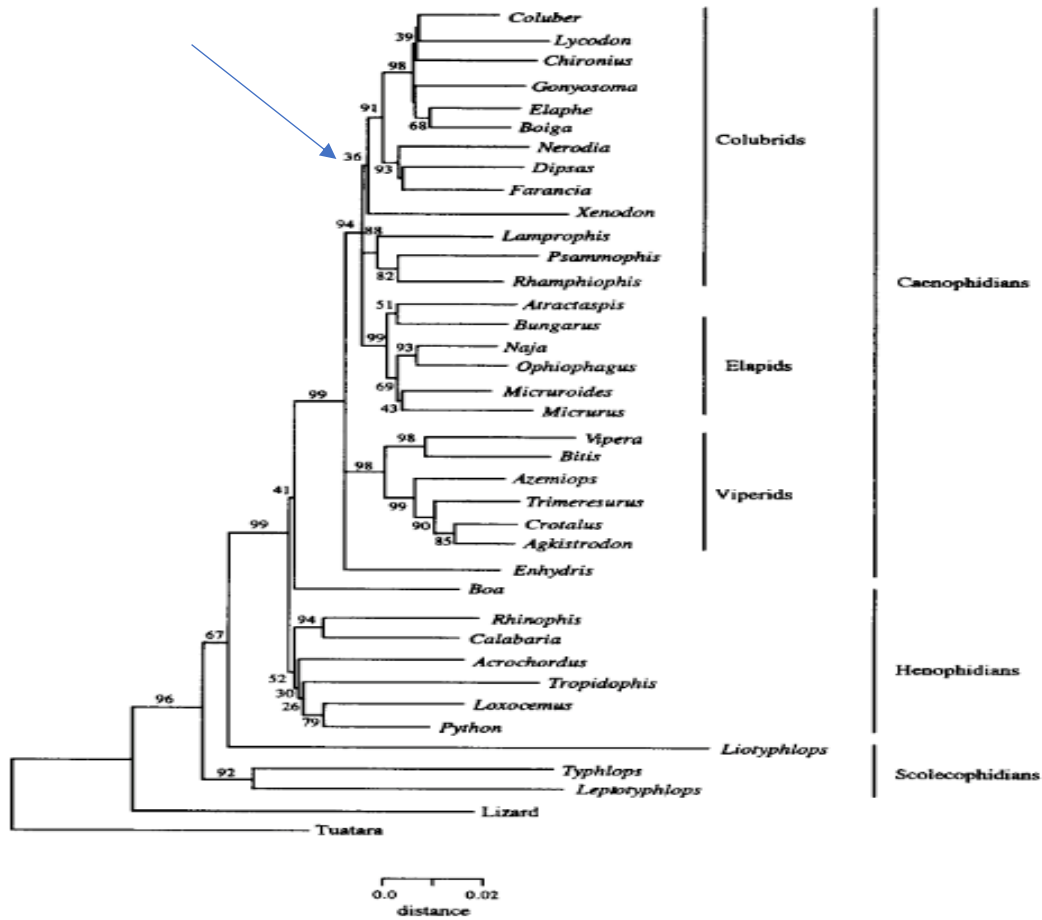
The conservation of synteny diagram in figure 5.16 was produced using the phylogeny between the DNA and amino acid sequences from all the snake species. This was also confirmed on a genomic level using ACT (Artemis Comparison Tool) whereby the genomic cluster sequences were blasted against each other and then plotted to show region of similarity (figure 5.15). The genes that proceed immediately upstream from CTSB gene in the cluster show a high conservation of synteny between species and as the cluster progresses upstream more individual characteristics become present. *C. v. viridis* only shows a smaller cluster than *T. elegans* and *N. naja* and shows that the cluster is not flanked by XPO1 like *T. elegans* and *N. naja* also. The blue region on *T. elegans* is showing a cluster duplication and to further investigate whether this was species specific a selection of beta-defensin genes from this cluster region were used to query if these were identified in other species. It was found that several beta-defensins in the region were specific to the genus *Thamnophis*. This was also confirmed in the genome comparison in figure 5.16 by the large green area in the *T. elegans* Beta-defensin cluster region.

The green regions highlighted in figure 5.16 show a paralogous set of genes that are only found within the *T. elegans* and *N. Naja* clusters and not the *C. v. viridis* sequence. Further investigation into these regions was carried out to establish whether they are a set of the beta-defensin genes found throughout more species of snakes and if they are specific to *Colubridae* and *Elapidae* families. The amino acid sequences of the second exon were chosen from a region in the *T. elegans* cluster to further look into this. TEBD41-43 were chosen (circled in figure 5.16) to BLAST against the snake sequences available in the whole shotgun sequencing database on NCBI. Table 5.5 shows which species had matches of over 50% identity with the genes blasted and what family they fall under. These genes were only found in *Columbidae* and *Elapidae* species and therefore it is possible that this region evolved after the last common ancestor after the split from the *Viperidae* family. Figure 5.14 taken from Heise *et al.* (1995) shows the phylogeny of snakes and where possibly the event of the rise of this of cluster region. Therefore, there is a need for further analysis to understand the mechanisms around this in more detail.

	TEBD27	TEBD29	TEBD30	TEBD32	TEBD33	TEBD41	TEBD42	TEBD43
<b>ELAPIDAE</b>								
<i>Notechis scutatus</i>						√	√	
<i>Laticauda colubrina</i>						√	√	
<i>Pseudonaja textilis</i>						√	√	√
<i>Hydrophis hardwickii</i>						√	√	
<i>Hydrophis melanocephalus</i>						√	√	
<i>Hydrophis cyanocinctus</i>						√	√	
<i>Naja</i>						√	√	√
<i>Ophiophagus hannah</i>							√	√
<i>Emydocephalus ijimae</i>						√	√	
<i>Laticauda laticaudata</i>						√	√	
<b>COLUBRIDAE</b>								
<i>Thamnophis sirtalis</i>	√	√		√	√	√	√	√
<i>Thamnophis elegans</i>	√	√	√	√	√	√	√	√
<i>Ptyas mucosa</i>						√	√	
<i>Pantherophis guttatus</i>						√	√	√
<i>Pantherophis obsoletus</i>						√	√	√
<i>Thermophis baileyi</i>								√

**Table 5.5 Genes common in Elapidae and Colubridae species.**

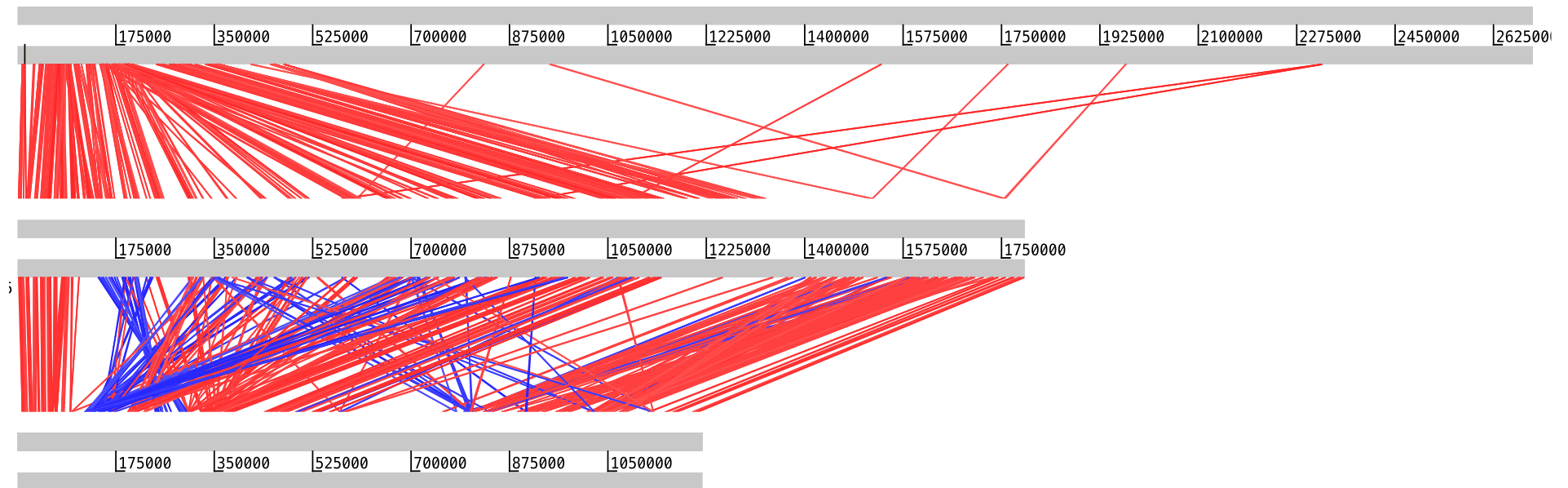
Blue boxes showing genes that are specific to *Thamnophis* genus. Identified by using tBLASTn of the amino acid sequences against Whole Genome Shotgun sequences with an % identity more than 50%.



**Figure 5.14** Phylogeny of snakes.

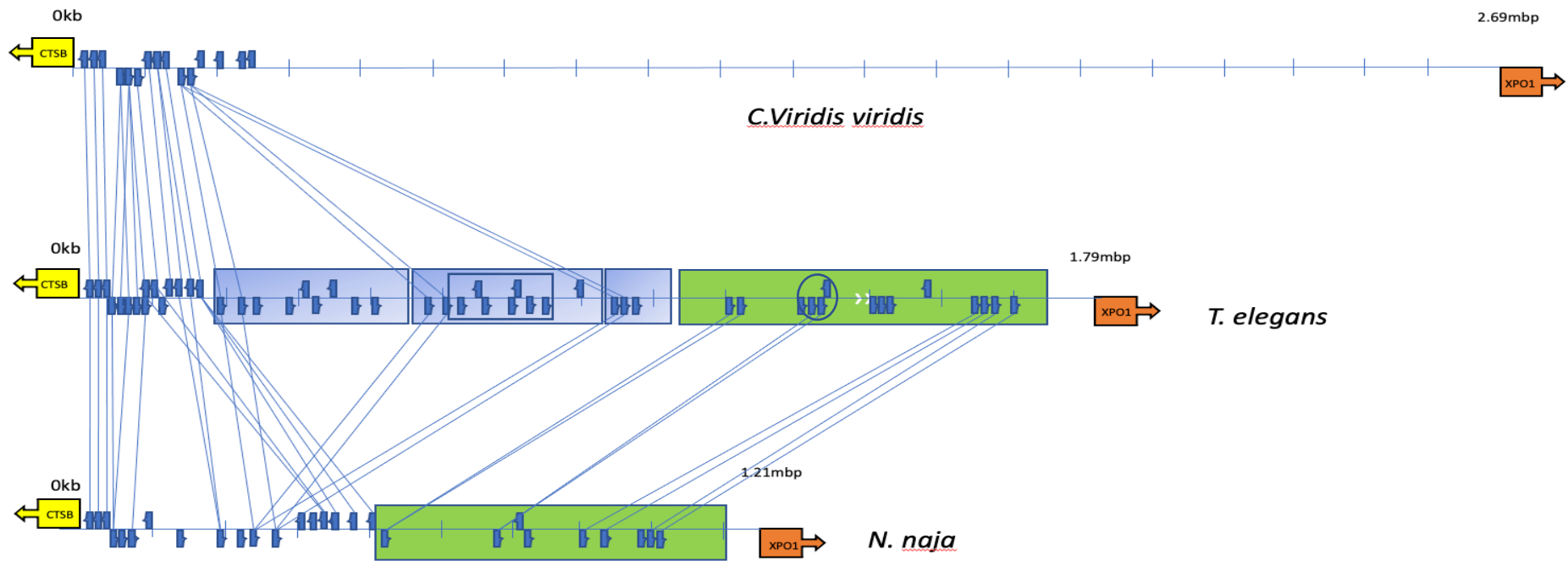
Where potential evolutionary event of this section of cluster (blue boxes in Figure 5.16) may have arisen (blue arrow). Taken from Heise et al. (1995)





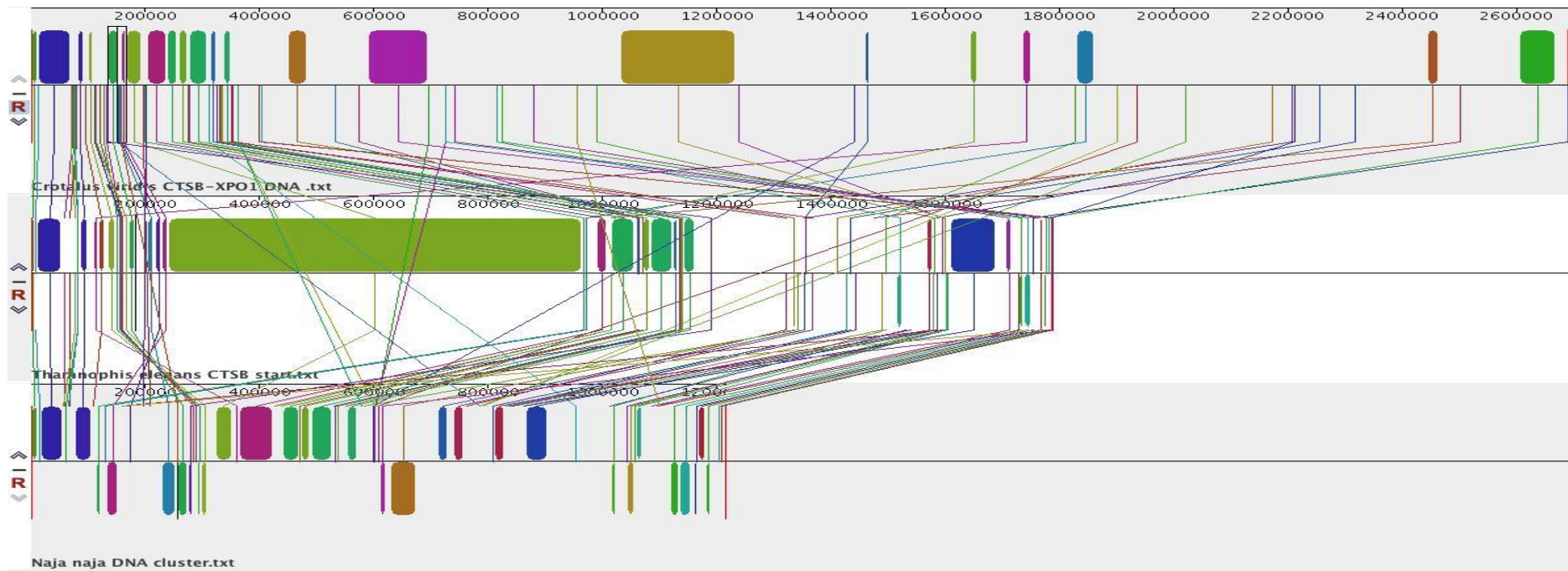
**Figure 5.15 Blast alignment of cluster region sequences taken from between CTSB and XPO1.**

Red line represents forward matches and blue represents reverse strand match of area of similarity. Top cluster is *Crotalus viridis viridis*, middle *Thanmophis elegans* and bottom is *Naja naja*. This diagram was produced in Artemis Comparison Tool (Carver et al. 2005)



**Figure 5.16 Conservation of synteny between snake clusters.**

Top cluster is *Crotalus viridis viridis*, middle if *Thamnophis elegans* and bottom is *Naja naja*. Genes are represented by dark blue boxes on the scale line. The blue shaded boxes represent cluster duplications as described in *T. elegans* and the green shaded boxes represents large area of conservation specific to *T. elegans* and *N. naja*. The circle represents genes taken for BLAST analysis to confer uniqueness of region to colubrid and elapid species. Blue Square represents blasted genes used to show beta-defensins specific to genus *Thamnophis*.



**Figure 5.17 Multiple cluster region alignment of DNA sequences from CTBS to XPO1 in snakes clusters.**

Large green region that may have expanded in Thamnophis genus. Other matching colours show areas of similarity. Built using Mauve Multiple Genome Alignment software (Darling et al. 2004).

## **5.5 Summary**

Snakes show a varied and differing variety and similarity of beta-defensins genes within their clusters. Each cluster was identified between the CTSB gene and XPO1 gene. The beta-defensins identified followed the typical 2 exon structure with a signal peptide associated with exon one and a mature peptide being tied to exon 2. The physical properties identified also showed a huge range of different charges which could play yet unknown roles.

Evolutionary analyses showed that the genes are undergoing purifying selection on a gene wide level which was shown in conservation of the first exon/signal peptide but in a more codon-based way showed regions between the highly conserved cysteine motif undergoing positive selection which may give rise to light changes in the tertiary structure. Mechanisms which may give rise to how the cluster undergoes expansion, and duplications within these, were identified including whole cluster duplications. This may allow further diversification of the beta-defensin genes within the cluster.

The cluster region containing the first beta-defensins immediately upstream of CTSB gene also show a high level of homology and synteny but as the genes moved downstream from this region, they became more diverse and family/species specific. It was identified that a region toward the end of the cluster was specific to elapidae and colubridae snake species and even more specific genes were identified that were gene specific.

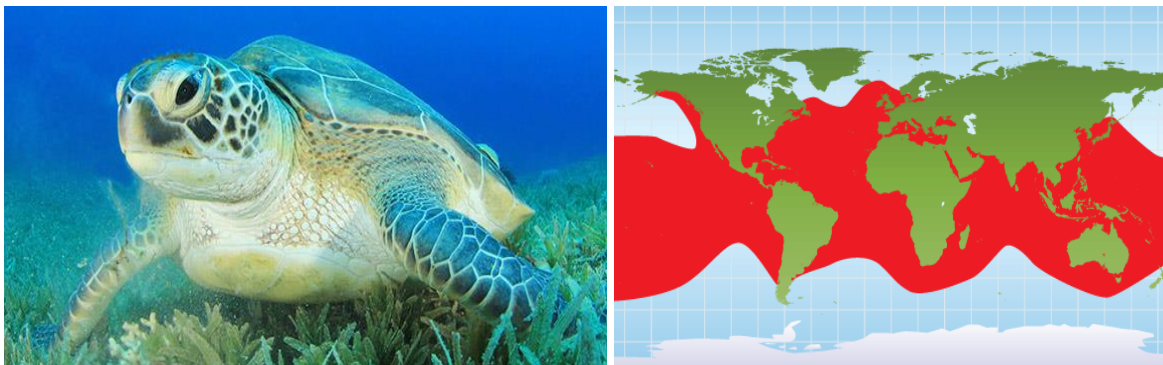
## Chapter 6 – TESTUDINES

### **6. Aims**

This chapter will focus on two testudines, one turtle and one tortoise. The numbers of genes present, genomic organisation and physical properties explored along with analysing conservation of synteny and describing similarities and differences within the cluster as a whole.

The two species are the Green Sea Turtle, *Chelonia mydas*, and The Goode's Thornscrub Tortoise, *Gopherus evgoodei*. The sequences described in this section were obtained by *In silico* means by the methods outlined in chapter 2. The beta-defensin genes were annotated based on the six-cysteine motif to establish the complete clusters. These will be described within each species and then compared between species and will try to answer questions that will describe possible reasons for variabilities within the gene of the cluster and how the cluster may have formed.

### **6.1 *Chelonia mydas* – Green Sea Turtle**



Photos: Turtle - <https://www.nwf.org/Educational-Resources/Wildlife-Guide/Reptiles/Sea-Turtles/Green-Sea-Turtle>  
Map - <https://greenseaturtlesendangered.weebly.com/habitat.html>

The Green Sea Turtle is a species of sea turtle of the family Cheloniidae and is the only species in the genus *Chelonia*. It has a worldwide range that extends throughout the tropical and subtropical seas in which there are subpopulations, the Atlantic and Pacific. They have nesting locations in approximately 80 countries. It is identified by its smooth heart-shaped

carapace which covers most of its body apart from its head and large front flippers. They can grow up to 3-4 feet in length and can weight up to 150kg. There are different colourings and markings in the subpopulations, but the main colour of the carapace is olive to black. These animals migrate long distances between feeding grounds where they feed on seagrasses as they have an herbivorous diet. Along with migrating long distances to feed they also migrate to nest and lay their eggs on beaches where they crawl onto the beach and bury their eggs in a nest, then hatchlings crawl to the water. In the wild the green sea turtle can live up to 80 years. The World Conservation Union (WCU) has classified the green sea turtle as endangered due to many causes including habitat loss, pollution, and hunting. It is illegal to collect, harm or kill them.

### **6.1.1 Data mining and cluster assembly**

The genomic sequencing data that was used for this analysis was obtained from The National Centre for Biotechnology Information (NCBI) genome assembly database. The genome was chosen as it was at chromosome level of genome assembly to allow the full cluster to be determined. The GenBank assembly accession number is GCA\_015237465.1 and was submitted to the database on 05/11/2020. As a starting point to narrow down the search of beta-defensins within the genome, the amino acid sequences of Cathepsin B (CTSB) and Translocating chain-associated membrane protein 2 (TRAM2) were obtained from the NCBI website. It was unknown if the cluster region resided between these two genes, however this starting point was used with the publication of the Komodo Dragon beta-defensins (van Hoek 2019). In this paper van Hoek and colleagues stated that CTSB and TRAM2 flank the Beta-defensin clusters in birds, turtles, and Crocodylia. These amino acid gene sequences were downloaded and used as a query when searching for the cluster region within the genome. Using the amino acids sequences obtained through the methods development, a concatenation of the partial sequences produced from the scaffold of the beta-defensins discovered in the *Chrysemys picta bellii* were used as a query to search against the genome, The Basic Local Alignment Search Tool (BLAST) with the tBLASTn program was chosen in order to find the region of the genome between CTSB and TRAM2. The region of the potential cluster was determined and resided between positions 9527031-10602441 on chromosome 3, a total length of approximately 1.07 Mb – GenBank accession number NC\_051243.1. This region was

masked using the RepeatMasker.org server (Smit *et al.*) to remove the repeat sequences from the DNA sequence. The was then translated into a 6-frame output using EMBOSS Sixpack program on The European Bioinformatics Institute (EMBL-EBI) website and this was utilised to highlight potential matches.

A concatemer approach outlined in the methods was employed to query the region that was established. Several concatemers were used that had previously been obtained through previous chapters. The tBLASTn program was applied against this region and highlighted on 6-frame output. This process, however, does not acquire all the exons and therefore other approaches were employed. Gene finding programs GENSCAN (Burge and Karlin 1997) and FGENESH (Solovyev *et al.* 2006) were employed to search for putative exons that were not initially found with the BLAST approach. Finally, regions of more than 3000bp of the repeats determined in the RepeatMasker analysis but not in the vicinity of already resolved exons can then be searched to exclude all potential regions where Beta-defensin exons may reside.

Splice site prediction was finalised using the online server by the Berkeley Drosophila Genome Project (Reese *et al.* 1997) and amino translations were ascertained from the DNA sequences of potential exons. Finally, iterative searches were performed using the newly identified beta-defensins against the cluster region.

### **6.1.2 Cluster organisation and Beta-defensin sequences**

A total of 39 beta-defensins were identified within this region and were numbered according to the position on the chromosome starting from the nearest gene to CTSB, in this case CMBD. Relative positions and genomic organisation along the DNA region are depicted in figure 6.1. Positions of each exon and sizes are available in appendix 3.1.



**Figure 6.1 Genomic organisation of the *C. mydas* Beta-defensin cluster.**

Each vertical line represents 100kb along the chromosome. The blue blocks represent the full genes, and the arrows dictate their orientation. The size of the squares (genes) or the distance between them are only representative and are not proportional to their actual size in the genome.

All the beta-defensins identified show the classical structure and consist of two exons. Exon 1 encodes a conserved signal peptide followed by the second exon encoding the mature antimicrobial peptide. The conserved defensin motif is present with common 6 cysteine domain and a glycine located in the N-terminal region, two positions upstream from the second cysteine, while another glycine is placed two or three positions upstream from the fourth cysteine with the rest of the amino acids being less conserved but show similarities where the genes have recently duplicated. This is observed in the multiple sequence alignment showing conservation motif (figure 6.2).

### **6.1.3 Physical Properties**

Each beta-defensin gene identified in this genome possesses a conserved signal peptide, and this was confirmed by the Performed using SignalIP – 5.0 server (Almagro Armenteros *et al* 2019) (Appendix 3.3) There is a wide range of charges and some of the beta-defensins in this cluster are anionic although most of the beta-defensins are cationic (Appendix 3.2). One such defensin, CMBD20 has a charge of -7 but this is similar to what was found in Crocodylia (Tang *et al* 2018) in that it has a long anionic pro-domain, and this may serve as a way to balance the charge of the defensin before undergoing further post translational modifications to produce the active mature peptide. Table 6.1 shows the charges between the long pro-domain beta-defensins minus the signal sequences and then the charge of the 2<sup>nd</sup> exon, which may closely represent the mature active form.



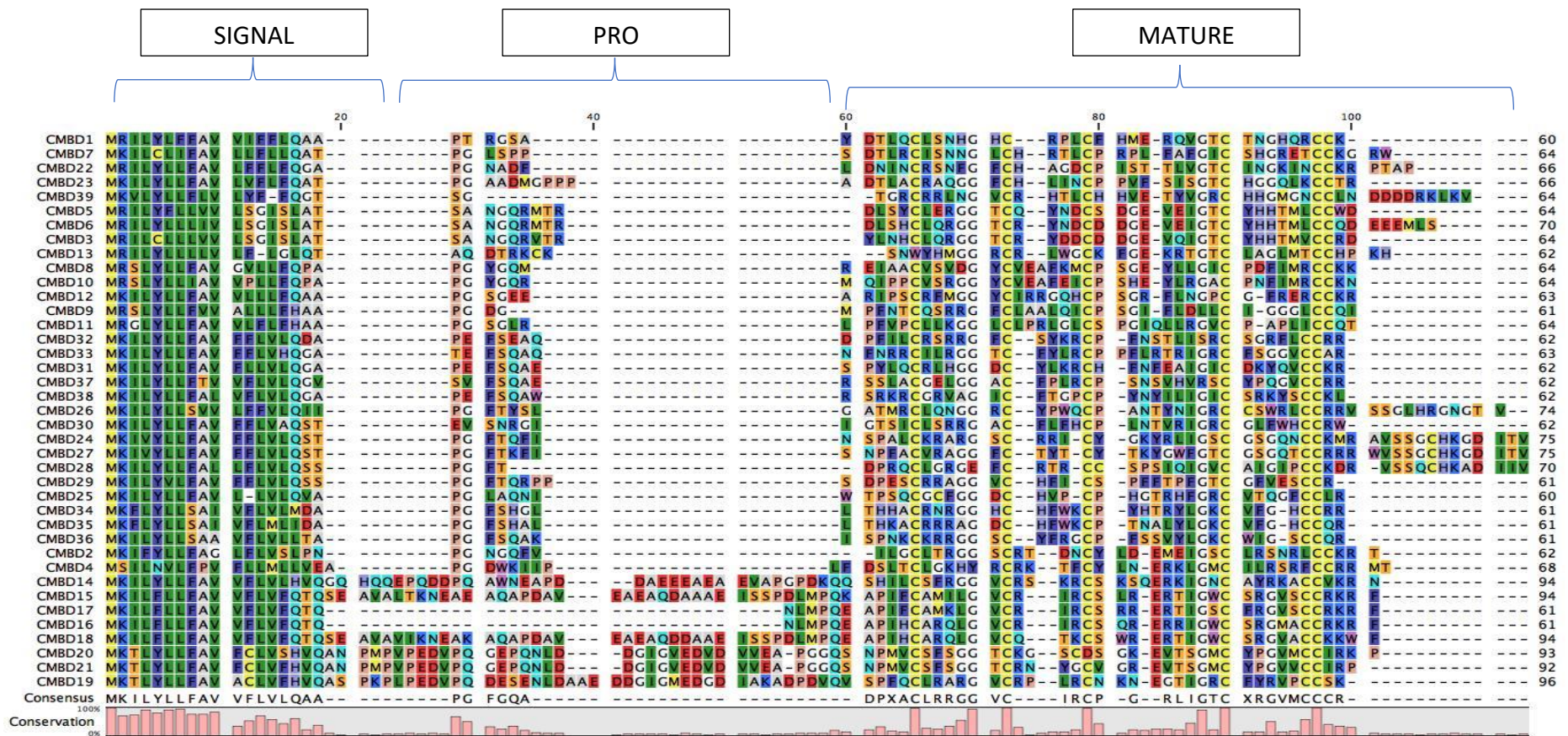
	Long pro-domain/mature peptide			Second Exon		
	pI	Net Charge	Mr	pI	Net Charge	Mr
CMBD14	6.12	-1	8444	10	10	5030
CMBD15	7.75	1	7941	10.21	8	4846
CMBD18	5.73	-1	8004	9.18	5	4868
CMBD19	4.54	-5	8424	9.18	5	4773
CMBD20	4.02	-7	7604	8.46	2	4348
CMBD21	3.93	-7	7587	8.47	2	4331

**Table 6.1 Charge differences between the longer pro-domain/mature peptides and the second exon in *C. mydas*.**

*Isoelectric point and molecular mass included.*

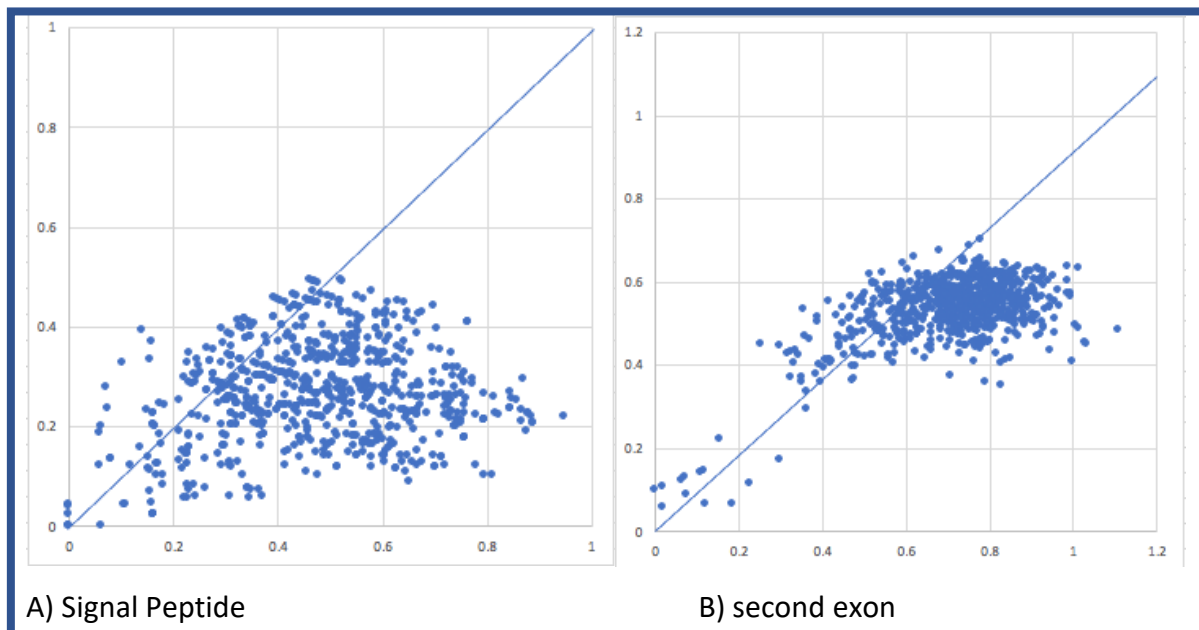
#### **6.1.4 Selection analyses**

Multiple sequence alignments were produced in CLUSTALX (Larkin *et al* 2007) and Codon alignments subsequently produced using the PAL2NAL server (Suyama *et al* 2006). These codon alignments were the used in pairwise comparisons between nucleotide sequences, the number of synonymous substitutions per synonymous site ( $dS$ ) and the number of nonsynonymous substitutions per nonsynonymous site ( $dN$ ) were estimated using the SNAP v.2.1.1 program at <http://www.hiv.lanl.gov> which implements Nei and Gojobori (1986) method (Korber 2000). The proportion of observed synonymous and nonsynonymous substitutions were plotted against each other (figure 6.3). Viewing the distributions between the signal peptide and the second exon there are slight differences on the distribution of the points. The signal peptide shows a greater degree of points distributed towards synonymous substitutions showing a high level of conservation between codons across the gene implying that is undergoing possible purifying selection pressures. However, the second exon shows that the distribution is closer to  $dS=dN$  but still showing a slight purifying selection. This is most likely down to the number of paralogues within the cluster having homology. When observing the second exons within the whole cluster you may expect there to be a greater degree of nonsynonymous substitutions due the variation of amino acid sequences present, therefore a site-wise analysis was performed to gain a better picture of the evolutionary dynamics within the individual sites within the gene.



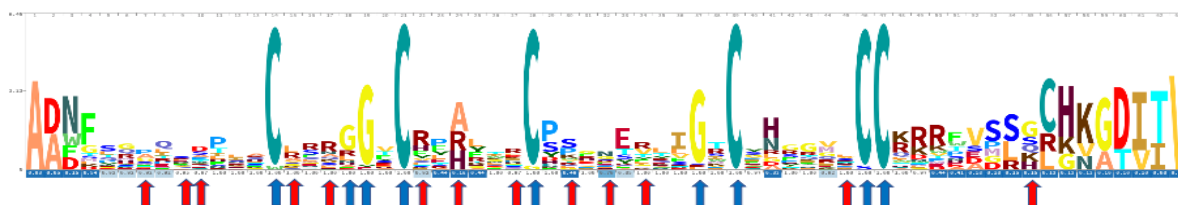
**Figure 6.2 Multiple sequence alignment of *C. mydas* beta-defensins cluster.**

Produced in Clustal X. Conservation of amino acids is shown in the legend underneath and show the typical conserved signal peptide at the start of the gene and in the mature peptide showing the 6 conserved cysteine residues along with glycine residues. Signal, Pro-peptide and Mature regions are also shown by the parentheses.



**Figure 6.3 Ratio of synonymous and nonsynonymous substitutions in *C. mydas*.**

Ratios within the signal peptide (A) and the second exon peptide (B) are nonsynonymous ( $d_N$ ) on the x axis and synonymous ( $d_S$ ) on the y axis. The diagonal lines represent  $d_N = d_S$  and is given for estimating selection pressures; dots above and below this line represent positive and purifying selection, respectively. Add to other figures



**Figure 6.4 Amino acid sequence Logo of second exons in *C. mydas*.**

Sites which are undergoing positive selection (red arrow) by one of more tests and purifying selection (blue arrow) tested by MEME, FEL and FUBAR in HYPHY. Logo produced on Skyline.org.

Within the second exon there are 13 positions that are undergoing positive selection and 9 residues undergoing negative/purifying selection. The positive-selection positions are located between the negative-selection positions. The negatively pressured amino acids are shown to be residues common to beta-defensins. These are the 6 cysteines that make up the covalent bonding that is seen throughout the defensin class along with the glycine residues

notably the GxC residues and the second and fourth cystine residues that make up the beta sheets integral to its structure (Tu *et al* 2015). However, the residues that are undergoing positive selection are located in the regions that contribute to the bends around these beta sheets and sited on the outside.

#### **6.1.5 Repeat Sequence landscape**

Repeat masker was performed using query species database set to tetrapod. The *C. mydas* defensin cluster region had over all 53.35% bases masked with the predominant repeat elements being retroelements at 70% of bases masked. LINES were around 53% of the retroelements and CR1 LINE being the most abundant at 74.2% of the LINES present. LTR elements accounted for about 43.5% of the retroelements. Around 27% of the repeat sequences were DNA transposons with hobo-Activator and Tourist/harbinger being the most abundant (table 6.2).

```

=====
file name: Chelonia_Mydas_DNA_Cluster.fa
sequences: 1
total length: 1075411 bp (1075411 bp excl N/X-runs)
GC level: 45.35 %
bases masked: 573743 bp ( 53.35 %)
=====

```

	number of elements*	length occupied	percentage of sequence
Retroelements	921	403121 bp	37.49 %
SINEs:	114	12704 bp	1.18 %
Penelope	167	43468 bp	4.04 %
LINEs:	522	214701 bp	19.96 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	324	159325 bp	14.82 %
R1/LOA/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	12	3506 bp	0.33 %
RTE/Bov-B	6	1259 bp	0.12 %
L1/CIN4	7	3131 bp	0.29 %
LTR elements:	285	175716 bp	16.34 %
BEL/Pao	0	0 bp	0.00 %
Ty1/Copia	0	0 bp	0.00 %
Gypsy/DIRS1	184	113631 bp	10.57 %
Retroviral	70	48437 bp	4.50 %
DNA transposons	799	155767 bp	14.48 %
hobo-Activator	333	61440 bp	5.71 %
Tc1-IS630-Pogo	25	4736 bp	0.44 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	290	48034 bp	4.47 %
Other (Mirage, P-element, Transib)	0	0 bp	0.00 %
Rolling-circles	2	104 bp	0.01 %
Unclassified:	53	9800 bp	0.91 %
Total interspersed repeats:		568688 bp	52.88 %
Small RNA:	3	183 bp	0.02 %
Satellites:	2	139 bp	0.01 %
Simple repeats:	112	4199 bp	0.39 %
Low complexity:	11	613 bp	0.06 %

\* most repeats fragmented by insertions or deletions  
have been counted as one element

The query species was assumed to be tetrapods  
RepeatMasker version 4.1.2-p1 , default mode

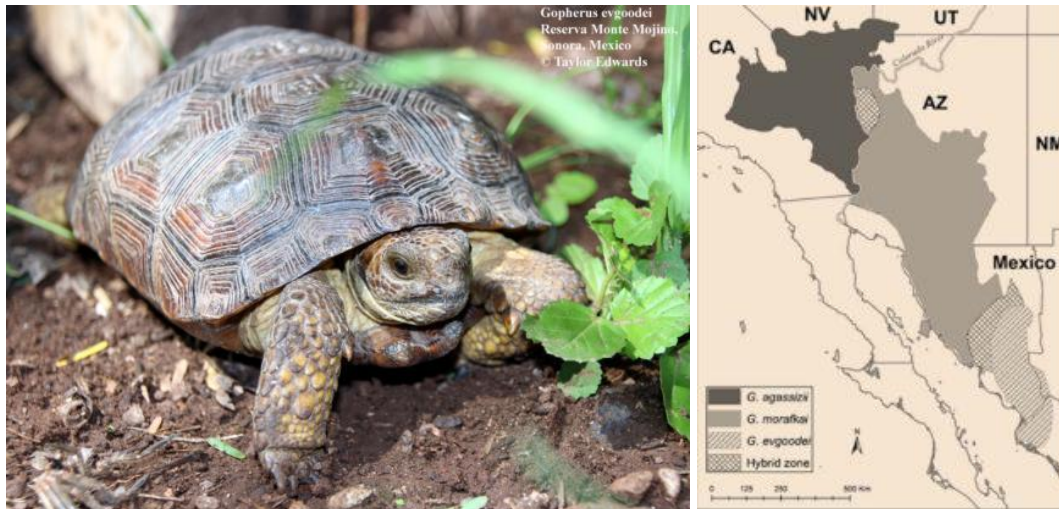
run with rmblastn version 2.2.27+  
FamDB: CONS-Dfam\_withRBRM 3.3

**Table 6.2 Repeat masker summary in *C. mydas***

*Displaying the different repeat sequences within the C. mydas cluster region. The tetrapod database was used as a reference for the repeat sequence matches in RepeatMasker program.*



## **6.2 *Gopherus evgoodei* - Goode's Thornscrub or Sinaloa Desert Tortoise**



Photos: Tortoise - <https://reptile-database.reptarium.cz/species?genus=Gopherus&species=evgoodei>

Map - <https://tucsonherpsociety.org/projects/mexican-tortoise-project/>

The Goode's Thornscrub Tortoise is a tortoise species from the *Testudinidae* family and is a member of 6 species of the genus *Gopherus* or Gopher Tortoise. As the alternative name suggests the Goode's Thornscrub tortoise's distribution is from the northern Sinaloa desert region of Mexico. It was first described in 2016 and is named after naturalist Eric V. Goode (Edwards *et al.* 2016). Their morphological features that make it distinct from their sister species are that they have a noticeably flatter shell profile with a shallower plastron, rounded footpads, multiple spurs on their radial-humeral joints and an orange tone on their skins (Edwards *et al.* 2016). Goode's Thornscrub Tortoise inhabits hills and low mountains with at least some boulders and rocky outcrops where it will burrow underneath these and where these aren't present it will dig burrows in soil and will use several of these a year (Edwards *et al.* 2016). Little more is known about daily activity, reproduction and movements this newly discovered species.

### **6.2.1 Data mining and cluster assembly**

As with the Green Sea Turtle the genomic sequencing data utilised for this analysis was obtained from The National Centre for Biotechnology Information (NCBI) genome assembly database. The genome was chosen as it was at chromosome level of genome assembly to

allow the full cluster to be determined. The GenBank assembly accession number is GCA\_007399415.1 and was submitted to the database on 25/07/2019. As a starting point to narrow down the search of beta-defensins within the genome, the amino acid sequences of CTSB and TRAM2 were obtained from the NCBI website. Orthologues were listed and the *G. evgoodei* sequences were downloaded. This starting point was used with the publication of the Komodo Dragon beta-defensins (van Hoek 2019). Using the amino acids sequences as a query to search against the genome, the tBLASTn program was chosen in order to find the region of the genome between CTSB and TRAM2. The region of the potential cluster was determined and resided between positions 18616371-19975803 on chromosome 3, a total length of approximately 1.35 Mb – GenBank accession number NC\_044324.1. This region was masked using the RepeatMasker.org server (Smit *et al.* 2006) to remove the repeat sequences from the DNA sequence. The was then translated into a 6-frame output using EMBOSS Sixpack program on The European Bioinformatics Institute (EMBL-EBI) website and this was utilised to highlight potential matches.

Unlike the approach of using concatemers outlined above, a slightly different methodology was employed. A FASTA file of the DNA coding sequences from the Green Sea Turtle were employed to run the search of the potential beta-defensins that reside within the region that was uncovered between CTSB and TRAM2. The DNA coding sequences from *C. mydas* were used as a query against the genome using the BLASTn program and due to the orthology of these sequences, the number of initial *G. evgoodei* sequences found was abundant. This DNA coding search approach did identify most of the genes present but in order to have confidence that all the genes were uncovered the concatemer approach was applied using the *C. mydas* amino acid sequences followed by the gene finding programs GENSCAN (Burge and Karlin 1997) and FGENESH (Solovyev *et al.* 2006) and finally searching the regions of more than 3000bp between repeat sequences shown by the running of the RepeatMasker program. Splice site prediction was finalised using the online server by the Berkeley Drosophila Genome Project (Reese *et al.* 1997) and amino translations were ascertained from the DNA sequences of potential exons. Finally, iterative searches were performed using the newly identified beta-defensins against the cluster region.

### 6.2.2 Cluster organisation and Beta-defensin sequences

A total of 47 beta-defensins were identified within this region and were numbered according to the position on the chromosome starting from the nearest gene to CTSB, in this case GEBD. Relative positions and genomic organisation along the DNA region are depicted in figure 6.5. Positions of each exon and sizes are available in appendix 3.4.



**Figure 6.5 Genomic organisation of the *G. evgoodei* beta-defensin cluster.**

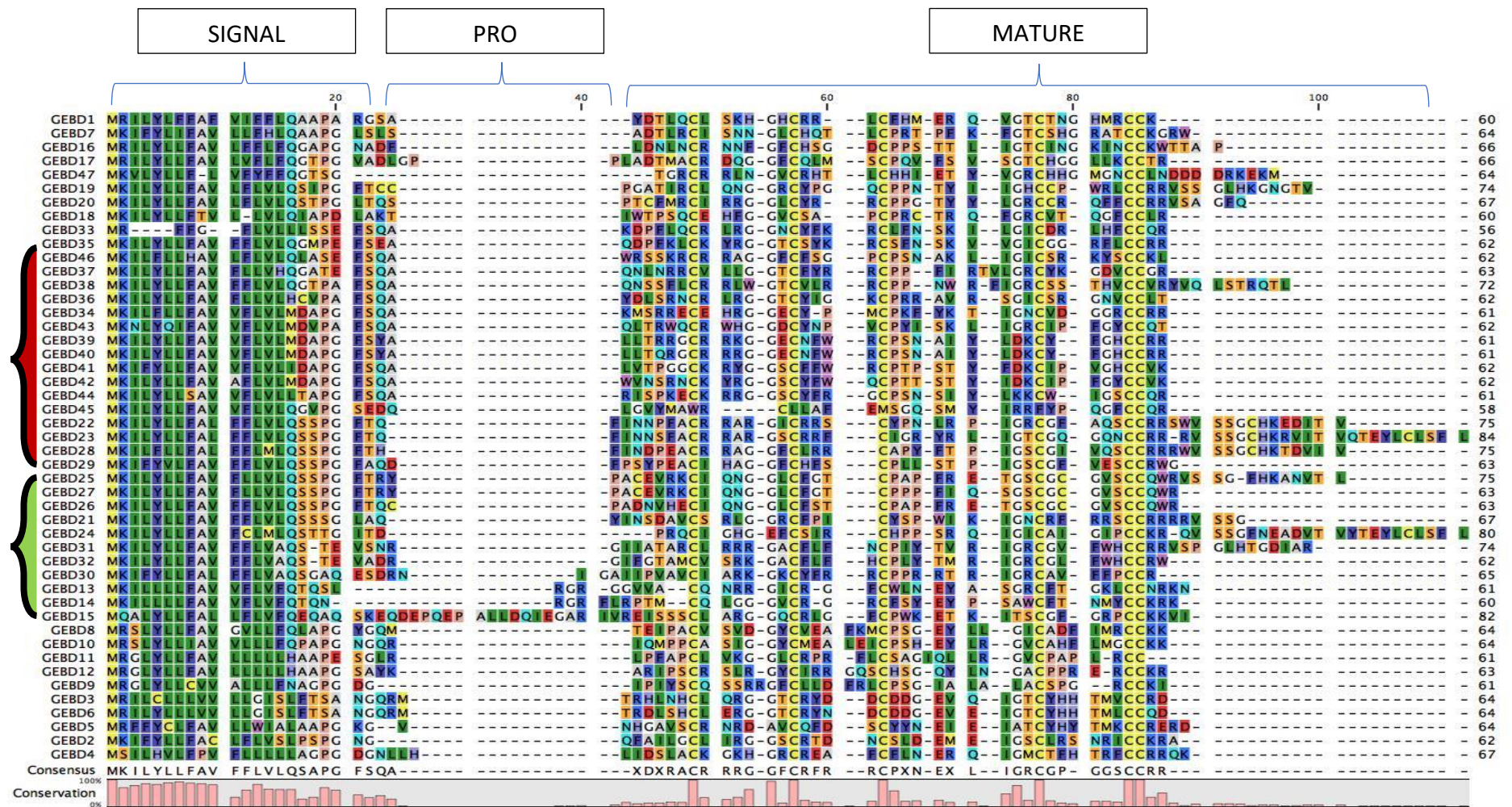
Each vertical line represents 100kb along the chromosome. The blue blocks represent the full genes, and the arrows dictate their orientation. The size of the squares (genes) or the distance between them are only representative and are not proportional to their actual size in the genome.

All the beta-defensins identified show the classical structure and consist of two exons. The conserved defensin motif is present with common 6 cysteine domain and a glycine located in the N-terminal region, two positions upstream from the second cysteine, while another glycine is placed two or three positions upstream from the fourth cysteine with the rest of the amino acids being less conserved but show similarities where the genes have recently duplicated. This is observed in the multiple sequence alignment showing conservation motif (figure 6.6).

### 6.2.3 Physical Properties

Each beta-defensin gene identified in this genome possesses a conserved signal peptide, and this was confirmed by the Performed using SignalIP – 5.0 server (Almagro Armenteros *et al* 2019) (appendix 3.6). There is a wide range of charges and some of the beta-defensins in this cluster are anionic although most of the beta-defensins are cationic (appendix 3.5). GEBD23 has a charge of +11 and GEBD6 has a charge of -6.





**Figure 6.6 Multiple sequence alignment of *G. evgoodei* beta-defensins cluster.**

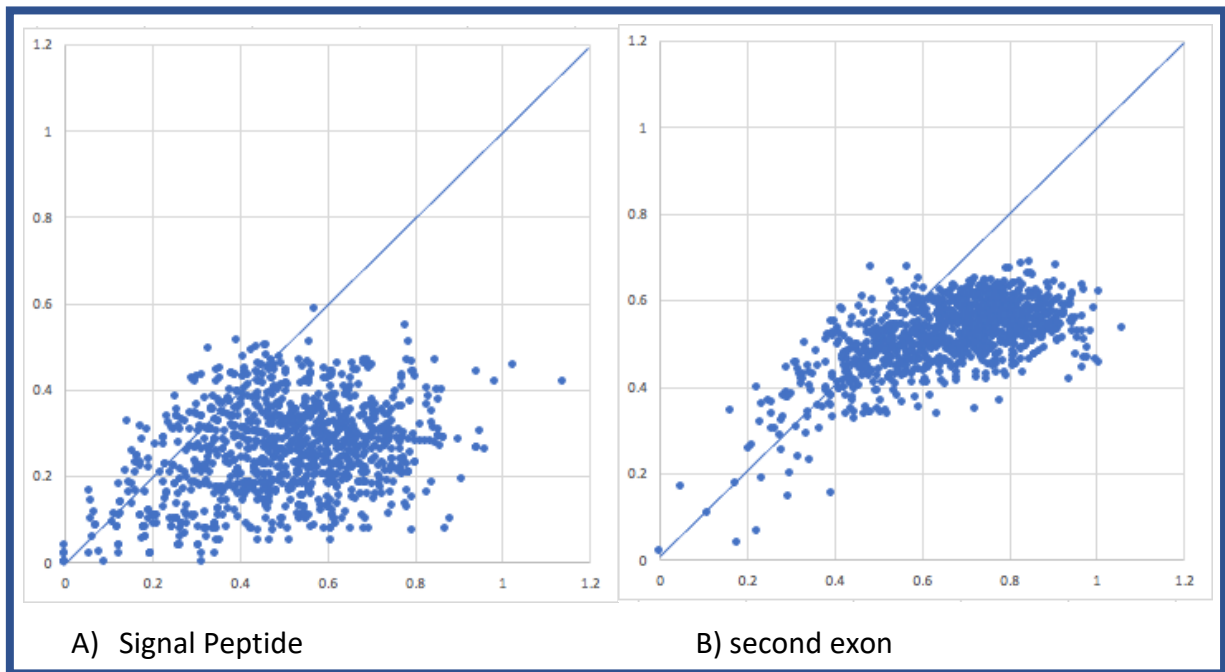
Produced in Clustal X. Conservation of amino acids is shown in the legend underneath and show the typical conserved signal peptide at the start of the gene and in the mature peptide showing the 6 conserved cysteine residues along with glycine residues. Signal, Pro-peptide and Mature regions are also shown by the parentheses.

#### **6.2.4 Selection Analyses**

Pairwise comparisons between nucleotide sequences were conducted whereby the number of synonymous substitutions per synonymous site ( $dS$ ) and the number of nonsynonymous substitutions per nonsynonymous site ( $dN$ ) were estimated using the SNAP v.2.1.1 program at <http://www.hiv.lanl.gov> which implements Nei and Gojobori (1986) method (Korber 2000). The proportion of observed synonymous and nonsynonymous substitutions were plotted against each other figure 6.7. Viewing the distributions between the signal peptide and the second exon there are slight differences on the distribution of the points. The signal peptide shows a greater degree of points distributed towards synonymous substitutions showing a high level of conservation between codons across the gene implying that it is undergoing possible purifying selection pressures. However, the second exon shows that the distribution is closer to  $dS=dN$  but still showing a slight purifying selection. This is most likely down to the number of paralogues having homology within the cluster. When observing the second exons within the whole cluster it is expected that there may be a greater degree of nonsynonymous substitutions due the variation of amino acid sequences present, therefore a site-wise analysis was performed to gain a better picture of the evolutionary dynamics within the individual sites within the gene. This was very similar to what was found in the Green Sea Turtle.

#### **6.2.5 Repeat Sequence landscape**

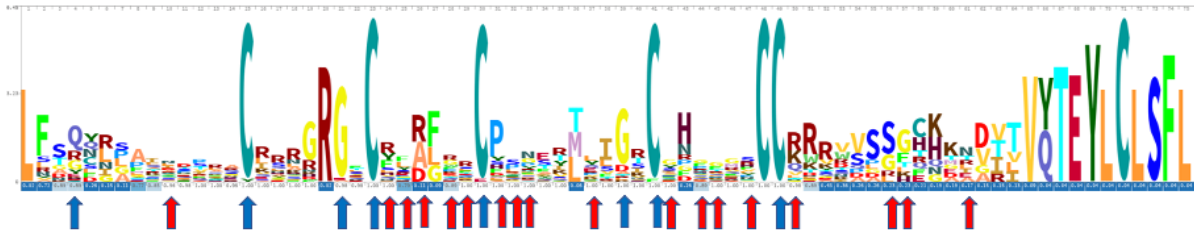
Repeat masker was performed using query species database set to tetrapod. The *G. evgoodei* defensin cluster region had over all 49.48% bases masked with the predominant repeat elements being retroelements at 64.7% of bases masked. LINES were around 56% of the retroelements and CR1 LINE being the most abundant at 79.9% of the LINES present. LTR elements accounted for 40% of the retroelements. Around 32% of the repeat sequences were DNA transposons with hobo-Activator and Tourist/harbinger being the most abundant (table 6.3).



**Figure 6.7 Ratio of synonymous and nonsynonymous substitutions in *G. evgoodei*.**

Ratios within the signal peptide (A) and the second exon peptide (B) show nonsynonymous ( $d_N$ ) on the y axis and synonymous ( $d_S$ ) on the x axis. The diagonal lines represent  $d_N = d_S$  and is given for estimating selection pressures; dots above and below this line represent positive and purifying selection, respectively.

Within the second exon there are 18 positions that are undergoing positive selection and 8 residues undergoing negative/purifying selection. The positive-selection positions are located between the negative -selection positions. The negatively pressured amino acids are shown to be residues that are common to beta-defensins. These are the 6 cysteines that make up the covalent bonding that is seen throughout the defensin class along with the glycine residues notably the GxC residues and the second and fourth cystine residues that make up the beta sheets integral to its structure (Tu *et al* 2015). However, the residues that are undergoing positive selection are in the regions that contribute to the bends around these beta sheets and sited on the outside surface of the peptide.



**Figure 6.8 Amino acid sequence Logo of the second exon in *G. evgoodei*.**

Sites which are undergoing positive selection (red arrow) by one of more tests and purifying selection (blue arrow) tested by FEL, FUBAR and MEME in HYPHY. Logo produced on Skyline.org

Phylogenetic analyses of the genes show an intraspecific gene clustering pattern, notably GEBD21-28 (highlighted in green parenthesis) and GEBD33-46 (highlighted in red parenthesis) in the multiple sequence alignment (figure 6.6). These regions also have traces of pseudogenes (data not shown) and these fit the ‘birth and death’ model of evolution first described by Nei and Hughes (1992). This can also be seen by the phylogenetic similarities shown by the genes highlighted in the tree (figure 6.9A). This model describes two main features a) an intraspecific gene clustering pattern and b) the presence of pseudogenes (Eirín-López *et al.* 2012). Also, noticeable in the data is comparing these regions of duplication with the corresponding dot-plot obtained by showing areas of similarity within the genomic region (figure 6.9B). Areas analogous to the multiple sequence alignment and phylogenetic tree are highlighted with green and red boxes on the dot plot and genomic organisation diagram (figure 6.9C). It can be observed that there are high regions of similarity which in turn translate to high regions of duplication of the beta-defensins within the cluster.



```

=====
file name: Gopher_Tortoise_DNA_cluster.fa
sequences: 1
total length: 1359433 bp (1359433 bp excl N/X-runs)
GC level: 44.06 %
bases masked: 672666 bp ( 49.48 %)
=====

```

	number of elements*	length occupied	percentage of sequence
Retroelements	1024	435369 bp	32.03 %
SINEs:	111	13907 bp	1.02 %
Penelope	168	43327 bp	3.19 %
LINEs:	549	244200 bp	17.96 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	356	195281 bp	14.36 %
R1/LOA/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	9	710 bp	0.05 %
RTE/Bov-B	8	2095 bp	0.15 %
L1/CIN4	3	1174 bp	0.09 %
LTR elements:	364	177262 bp	13.04 %
BEL/Pao	0	0 bp	0.00 %
Ty1/Copia	0	0 bp	0.00 %
Gypsy/DIRS1	279	135661 bp	9.98 %
Retroviral	61	30665 bp	2.26 %
DNA transposons	914	215652 bp	15.86 %
hobo-Activator	408	81748 bp	6.01 %
Tc1-IS630-Pogo	23	2866 bp	0.21 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	313	78179 bp	5.75 %
Other (Mirage, P-element, Transib)	0	0 bp	0.00 %
Rolling-circles	1	155 bp	0.01 %
Unclassified:	75	14175 bp	1.04 %
Total interspersed repeats:		665196 bp	48.93 %
Small RNA:	5	402 bp	0.03 %
Satellites:	2	128 bp	0.01 %
Simple repeats:	137	5673 bp	0.42 %
Low complexity:	27	1456 bp	0.11 %

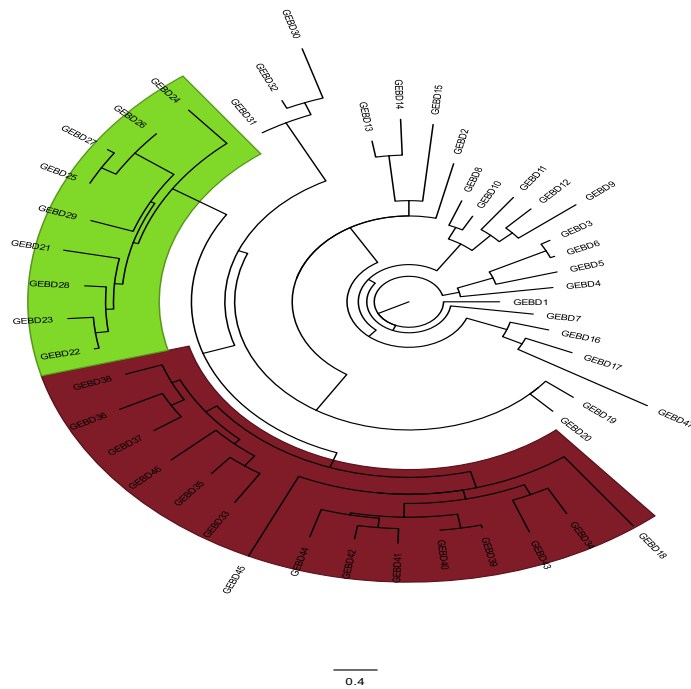
\* most repeats fragmented by insertions or deletions have been counted as one element

The query species was assumed to be tetrapods  
RepeatMasker version 4.1.2-p1 , default mode

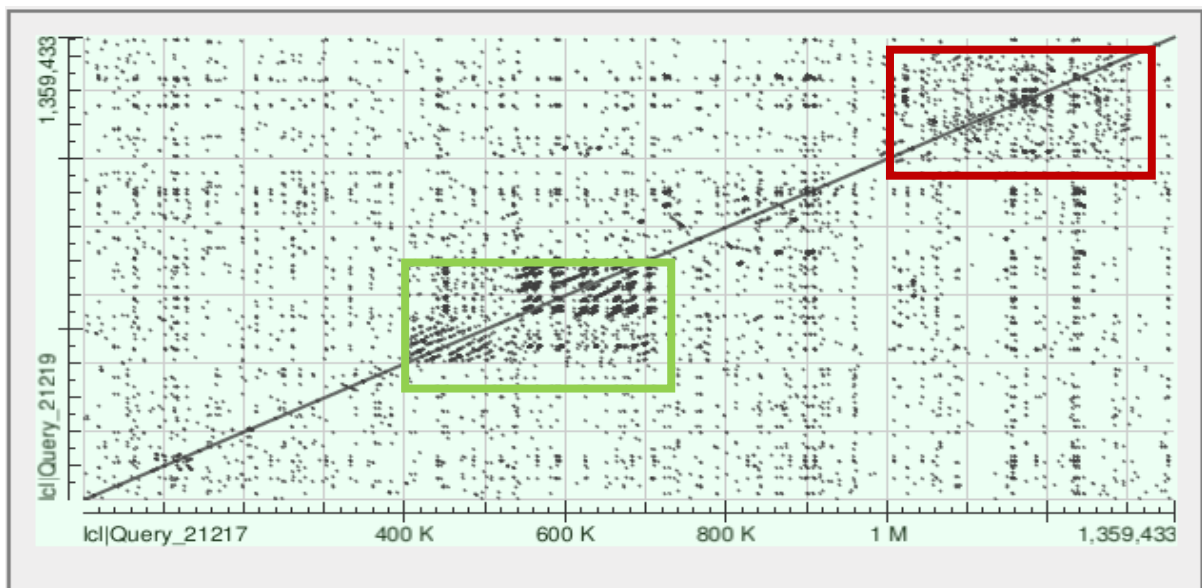
run with rmblastn version 2.2.27+  
FamDB: CONS-Dfam\_withRBRM\_3.3

**Table 6.3 Repeat masker summary for *G. evgoodei*.**

Displaying the different repeat sequences within the *C. v. viridis* cluster region. The tetrapod database was used as a reference for the repeat sequence matches in RepeatMasker program.



A) Phylogenetic tree of *G. evgoodei* beta-defensins. Corresponding genes highlighted showing gene of high similarity.

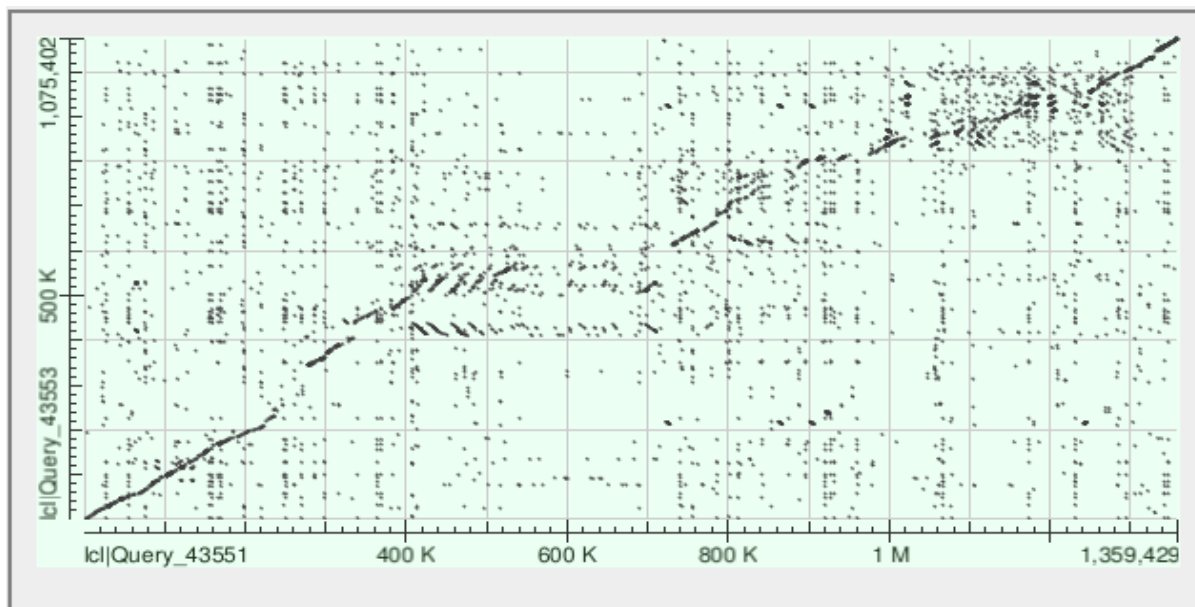


B) Dot plot of cluster region showing regions of high duplication.

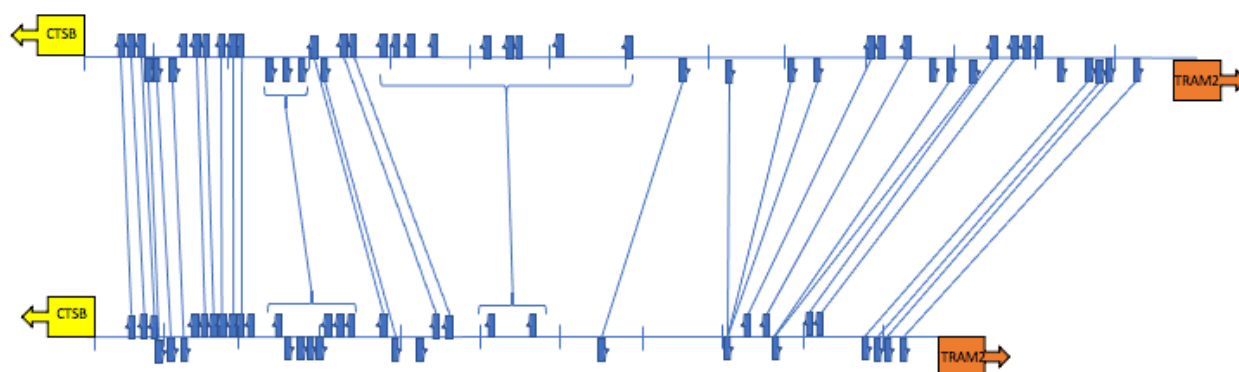


C) Genomic organisation with corresponding area of high duplication when compared to the dot plot and phylogenetic tree.

**Figure 6.9 Relationship of Genomic organisation, regions of high duplication and Phylogeny in *G. evgoodei*.**



A)



B)

**Figure 6.10 Conservation of Synteny between Testudine Cluster regions**

- A) Dot plot comparison of both *G. evgoodei* (x-axis) and *C. mydas* (y-axis). Areas of high homology can be observed by the diagonal line running upwards from left to right. Areas of duplication and expansion can be observed in areas between 400-700kb on the *G. evgoodei* cluster region.
- B) Genomic organisation of both *G. evgoodei* (top) and *C. mydas* (bottom). Lines represent paralogues of genes and parenthesis depicting where areas have undergone expansion and duplication.

### 6.3 Conservation of synteny

Between the two species there is an overall high level of conservation of synteny. Analysis shows that both species had CTSB as one flanking gene of the cluster and another being TRAM2. Similarly to the other species investigated in this work the region and genes closest to the start of the cluster showed greatest degree of homology. Moving downstream from CTSB differences started to emerge. In the *G. evgoodei* cluster there have been several expansions due to duplication events which have given a greater number of genes within the cluster when compared to the *C. mydas* cluster. A notable example can be seen in figure 6.10A where a number of duplications have occurred, however, these are recent as the level of similarity between the genes is high. There is also a region on the *C. mydas* cluster that has undergone a similar direction. Moving towards the end of the cluster it appears to revert to being homologous. One could hypothesise that these changes have occurred due to the different environments that each of these species inhabits and therefore been exposed to different challenges that these may exhibit.

### 6.4 Summary

The Gopher Tortoise and the Green Sea Turtle have a beta-defensin cluster that is located on chromosome three flanked by CTSB and TRAM2 genes. This is the same in birds and different to snakes and lizards. Testudines have a substantial array of beta-defensin gene number, ranging from 39 on the turtle to 47 in the tortoise. They both share similarities in their characteristics with both species possessing several genes that have a long propeptide in the primary amino acid sequences having a possible function in balancing the charges within the peptide. They have a conserved signal peptide and a more diverse mature peptide and within these possess a number of sites which are undergoing positive selection indicating that the species are still having challenges to pathogens which may be evolving equally in order to infect. There is evidence of the 'Birth and death' model for gene duplication. They had similar repeat sequence landscapes with nearly half the region dominated by these sequences. Conservation of synteny shows that they have very similar organisation except for regions where duplications have occurred in more recent evolutionary history.



## **Chapter 7 – MSBD1 expression and purification.**

### **7. Aims**

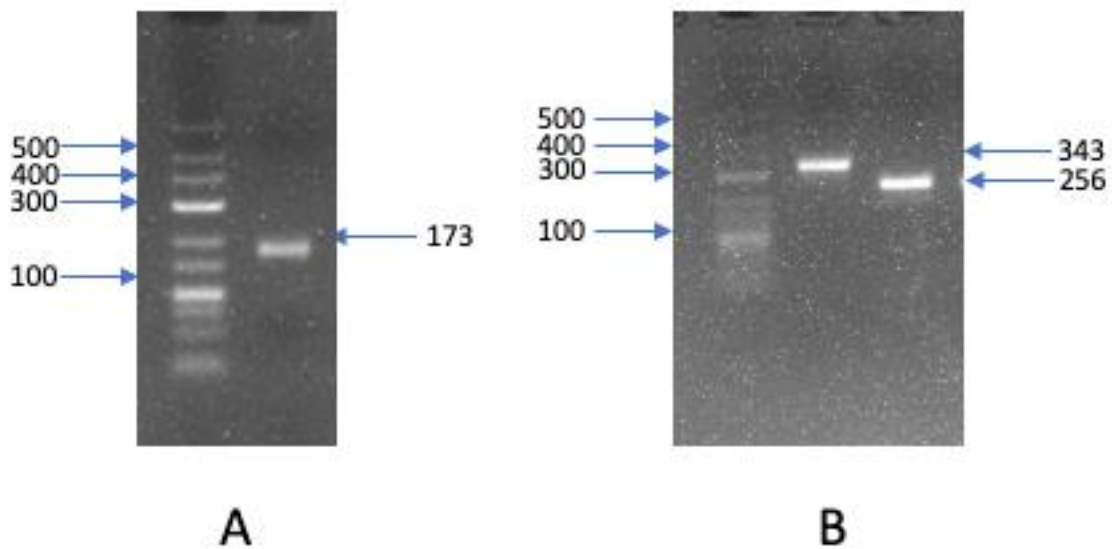
The aim of this project was to design a strategy to isolate, express and purify a mature Beta-defensin peptide in order to study its properties and establish antimicrobial activity. The gene under investigation was isolated from a carpet python (*Morelia spilota*) and was designated MSBD1.

### ***7.1 DNA extraction and RF cloning of MSBD1.***

After the DNA extraction process laid out in methods section 9.2.1. successful isolation of MSBD1 was achieved using the large hybrid primers which were designed for the restriction free cloning procedure. The estimated PCR fragment size was determined using the DNA sequence for a beta-defensin from the *Python bivittatus* genome (Genbank assembly accession number GCA\_000186305.2). Considering the primer length and second exon length the estimated size was 176bp (figure 7.1A). However, it should be noted that the fragment is isolated from a carpet python so could be of different size. This was sent for sequencing once it had been cloned into the plasmid. To clone this gene into the plasmid to create a fusion with the N- terminus of the Maltose Binding Protein a second PCR reaction was performed as outlined in methods section 9.2.3. Figure 7.1B depicts the amplification of the Multiple Cloning Site (MCS) to establish whether the clone had been successful. When this reaction was amplified alongside a control plasmid (without insert) it showed that the gene had been successfully inserted into the plasmid. The sizes and sequences of the products shown on the gel in figure 7.1 are described in figure 7.2.

As part of this process the methylated template parental plasmid (without MSBD gene insert) had to be digested using Dpn1 restriction enzyme (refer to methods section 9.2.3). As Dpn1 only digested methylated DNA only the unmethylated newly cloned plasmid would be transformed further downstream. It should be noted that an overnight digestion was needed for successful digestion as when the sample was sent for sequencing the parental plasmid was in such quantities that it was not fully digested. Therefore, this was showing as an unsuccessful clone even though when using PCR to probe for the insert showed that a

successful clone has in fact been achieved. This was due to using the hybrid primers instead of primer designed to amplify the MCS.



**Figure 7.1 Agarose Gel Electrophoresis of PCR gene product and Plasmid Clone**

A) Successful amplification of gene of interest from DNA isolated from *M. spilotota*. lane one shows DNA ladder and lane two showing amplified product of MSBD1 plus hybrid primer sequences 173bp in length. B) PCR amplification of the plasmid insert; lane 1 showing DNA ladder; lane 2 successful amplification of plasmid insert using primers to amplify the MCS; lane 3 MCS amplified on control plasmid without insert.

A) Isolated MSDB1 PCR insert

CGGGGAGAACCTGTACTTCCAG **AAGGGGGACCTTTATGACAGCCTAGTGTGCCACAACAATC**  
**ATGGACACTGCCGGAGACTGTGTTTCCACCGTGAACAGATAATCGGAACTTGCACCAATGGC**  
**CGGCAACGCTGCTGCAAATGA** GAATTCCTGCAGGTAATTAAATAAGCT 173bp

B) MCS without cloned MSBD1

GTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTAATTCGAGCTCGAACAACAAC  
AACATAACAATAACAACAACCT **CGGGGAGAACCTGTACTTCCAG** **ATGCTGATGGGCGGCCG**  
**CGATATCGTCGACGGATCC** GAATTCCTGCAGGTAATTAAATAAGCTTCAAATAAAACGAAA  
GGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTGTCG **GTGAACGCTCTCCTGA**  
**GTAGGACA** 256bp

C) Inserted sequence into MSC

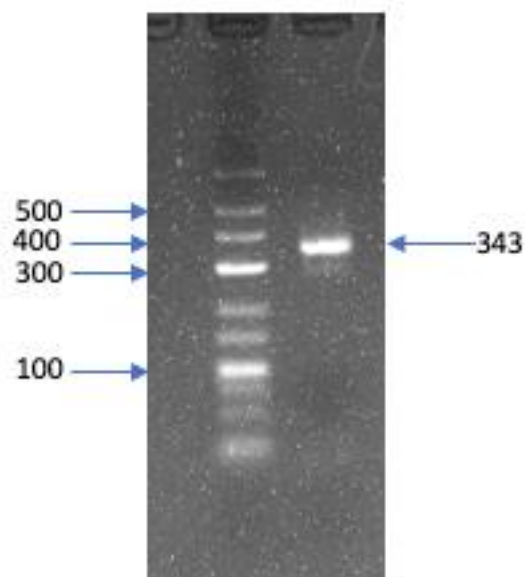
GTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTAATTCGAGCTCGAACAACAAC  
AACATAACAATAACAACAACCT **CGGGGAGAACCTGTACTTCCAG** **AAGGGGGACCTTTATGA**  
**CAGCCTAGTGTGCCACAACAATCATGGACACTGCCGGAGACTGTGTTTCCACCGTGAACAGA**  
**TAATCGGAACTTGCACCAATGGCCGG** **CAACGCTGCTGCAAATGA** GAATTCCTGCAGGTAAT  
**TAAATAAGCT**TCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGT  
TGTTTGTGCG **GTGAACGCTCTCCTGAGTAGGACA** 343bp

**Figure 7.2 DNA Sequences and sizing of gene insert and Plasmid MCS PCR products.**

*A) PCR amplified MSBD1 gene from extracted genomic DNA from the snakeskin. B) MCS sequence. Red highlighted sequence is removed from the original plasmid when amplified with the PCR insert. C) Amplified product used for sequencing showing inserted sequence into the plasmid. Blue lettering is the plasmid annealing sequence of one half of the hybrid primers and the red lettering is the target annealing sequence of the other half to target MSBD1 within the genomic DNA. Yellow highlighted sequence if MSBD1. Green lettering is primer pair used for sizing and confirming product has been inserted along as well as being used for sequencing the insertion. Sizes of products shown with the sequences.*

## 7.2 Electroporation and transformation.

Successful transformation into *E. coli* DH5 $\alpha$  through electroporation was achieved using the adapted method originally produced by Gonzales *et al* (2013). Figure 7.3 shows colony PCR on a selected successfully cloned colony with ampicillin for selection. *E. coli* DH5 $\alpha$  was used as a holding cell for the purposes of using as a stock for when more plasmid was needed and supply enough purified plasmid for sequencing. This plasmid was extracted using GeneJet plasmid miniprep kit (thermofisher) and was sent to The University of Birmingham Functional Genomics Laboratory. Successful cloning was achieved; therefore the plasmid was isolated and used to transform the plasmid into the SHuffle<sup>®</sup> (New England Biolabs) expression strain to allow for correct expression and folding of the protein including the ability to for the cysteine bonds needed for the MSBD1 beta-defensin by expressing cytoplasmic DsbC which provides isomerase activity for the correct formation and bonding within the protein.



**Figure 7.3 Colony PCR of inserted clone into the MCS.**

### ***7.3 Sequencing of plasmid insert and isolated MSBD1 gene.***

Through sequencing of the plasmid, the method employed to isolate the gene of interest, MSBD1, from the extracted genomic DNA and using a restriction free cloning method with the dual specific primer pairs was an efficient method to extract, isolate and clone a gene of interest into a plasmid. The primer design was done using genomic sequence from the Burmese Python assembly on NCBI, so it was quite fortunate that the isolate gene was in the same reading frame. The second exon was chosen as this would closely represent a mature beta-defensin peptide. The sequence and translation are as follows.

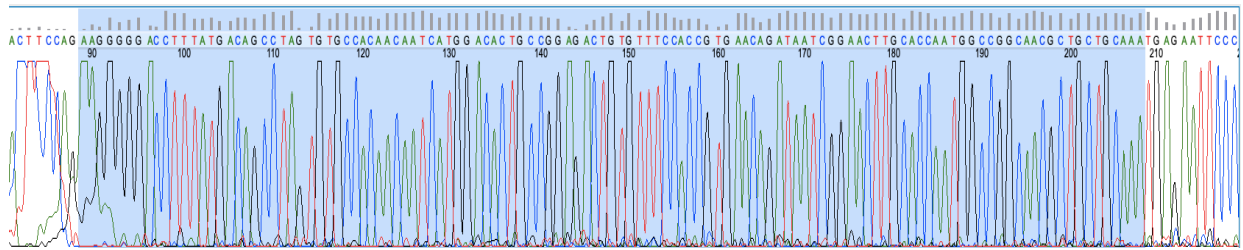
#### ***DNA SEQUENCE***

```
AAGGGGGACCTTTATGACAGCCTAGTGTGCCACAACAATCATGGACACTGCCGGAGACTGTGTTTCC  
ACCGTGAACAGATAATCGGAACTTGCACCAATGGCCGGCAACGCTGCTGCAAA
```

#### ***TRANSLATED SEQUENCE***

```
KGDLYDSLVCHNNHGHCRRLCFHREQIIGTCTNGRQRCK
```

Figure 7.4 depicts the electrophoretogram from the raw sequence showing the correct insertion position and reading frame to allow expression of the MBP-MSBD1 fusion protein. Although a lysine is position 1 of the cleavage site as well as the first amino acid of MSBD1 and contrary to the manufacturers recommendations it still allowed the peptide to be cleaved off with TEVp (highlighted in green in figure 7.5). In addition to this the advantage of using a restriction free cloning method allowed the sequence to be inserted with matched the genomic sequence without having overhangs as per the restriction methods for cloning. This gives a truer representation of the peptide.



**Figure 7.4 Sequence Electrophoretogram of MSBD1 insert.**

*Insertion of MSBD1 into the MCS of the plasmid (highlighted in blue). TEVp cleavage site is immediately upstream with N-terminal of maltose binding protein upstream of this cleavage sequence.*

```

1  G E P V L P E G G P L * Q P S V P Q Q S
1  G R T C T S R R G T F M T A * C A T T I
1  G E N L Y F Q K G D L Y D S L V C H N N
1  GGGGAGAACCTGTACTTCCAGAAAGGGGACCTTTATGACAGCCTAGTGTGCCACAACAAT
1      10      20      30      40      50
1  CCCCTCTTGGACATGAAGGTCTTCCCCCTGGAAATACTGTCCGATCACACGGTGTGTTGTTA
21  W T L P E T V F P P * T D N R N L H Q W
21  M D T A G D C V S T V N R * S E L A P M
21  H G H C R R L C F H R E Q I I G T C T N
61  CATGGACACTGCCGAGACTGTGTTTCCACCGTGAACAGATAATCGGAACTTGCACCAAT
61      70      80      90      100     110
61  GTACCTGTGACGGCCTCTGACACAAAGGTGGCACTTGTCTATTAGCCTTGAACGTGGTTA
41  P A T L L Q M R I P C R * L N K L Q I K
41  A G N A A A N E N S L Q V I K * A S N K
41  G R Q R C C K * E F P A G N * I S F K *
121  GGCCGGCAACGCTGCTGCAAATGAGAATTCCCTGCAGGTAATTAAATAAGCTTCAAATAA
121      130     140     150     160     170
121  CCGGCCGTTGCGACGACGTTTACTCTTAAGGGACGTCCATTAATTTATTCGAAGTTTATT
61  R K A Q S K D W A F R F I C C L S V N A
61  T K G S V E R L G L S F Y L L F V G E R
61  N E R L S R K T G P F V L S V V C R * T
181  AACGAAAGGCTCAGTCGAAAGACTGGGCCCTTTCGTTTTATCTGTTGTTTGTTCGGTGAACG

```

**Figure 7.5 Genetic translation map of sequenced region of plasmid.**

*Insertion of MSBD1 (Amino acid sequence highlighted in yellow). The cleavage recognition sequence for TEVp is highlighted in green and this also shows that correct reading frame for MSBD1 downstream and the blue arrow indicates the TEVp cleavage site. The sequences highlighted in red show the plasmid annealing section of the hybrid primers used for the restriction free cloning of the insertion MSBD1 sequence. Maltose Binding Protein sequence is upstream of the TEVp cleavage site.*

#### **7.4 Expression and solubility testing**

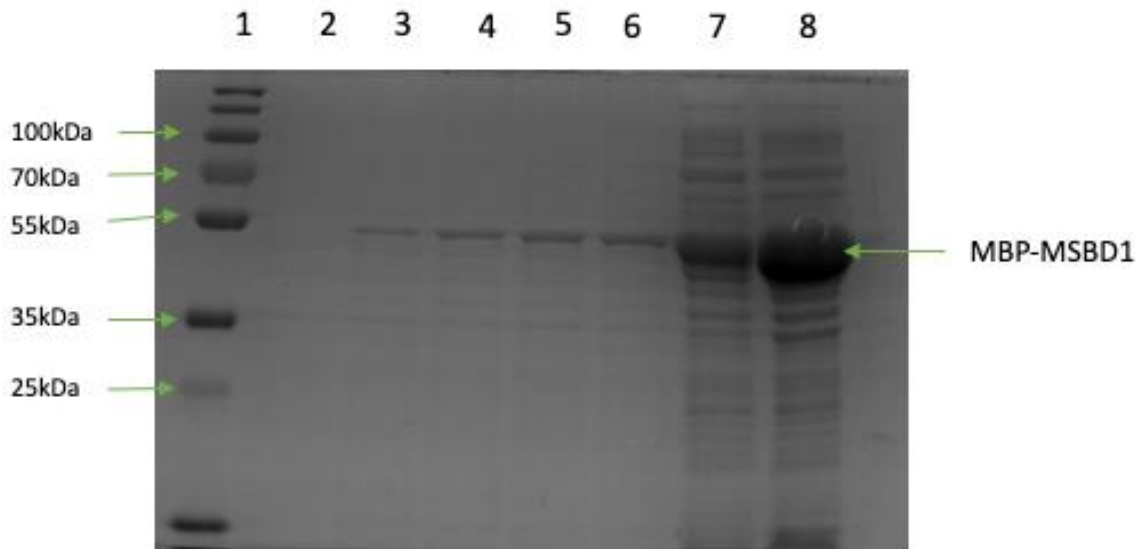
A 10ml culture of the transformed Shuffle® *E. coli* was grown overnight in LB broth media supplemented with 10mg/ml of ampicillin shaken at 37°C and this was used to inoculate 1L of LB broth with 10mg/ml of ampicillin shaken at 150rpm at 37°C. Expression of the MBP-MSBD1 fusion protein was achieved under the induction of 1mM IPTG for 4 hours once the OD<sub>600</sub> had reached 0.4. Figure 7.6 shows the expression profile of 0-4 hours (lanes 2-6). 1ml samples were taken and aliquoted into SDS-PAGE sample buffer for analysis to show expression was occurring. After 4 hours the cells were harvested by centrifugation for 30 minutes at 4000 x g. The pellet was immediately frozen at -20°C before lysing. Sonication was performed using the procedure set out in methods section 9.2.7. Once this had been conducted the mixture was once again centrifuged for 30 minutes at 4000 x g. The supernatant was drawn off and filtered through a 0.2µm membrane to allow for further downstream processing.

To find out whether the protein was being expressed as inclusion bodies, the solubility was tested once the cells had been harvested and lysed. A small aliquot was taken from the post sonication mixture and was centrifuged to pellet the cell debris. After centrifugation samples of the soluble and insoluble fractions were separated by SDS-PAGE and the cell pellet was resuspended in ddH<sub>2</sub>O. 20µl of each sample was added to SDS-PAGE sample buffer and heated for 10mins at 95°C to establish whether the protein was soluble or not. It can be seen in Figure 7.6 lanes 7 & 8. Lane 7 is of the cell pellet and lane 8 of the supernatant. Much of the protein is soluble, which is a characteristic of MBP, although some can be seen in the cell pellet. This could be because the lysing procedure was not sufficient or that some of the protein is misfolded and causing aggregates and therefore not soluble. It was noted in Li & Yeong (2011) that when producing a beta-defensin fusion protein with MBP caused a soluble aggregate to be produced so this is what can be seen in the SDS-PAGE gel.

To investigate at this stage as to whether the supernatant pre-affinity chromatography purification and before removal of the MBP affinity tag could be used as a screening for antimicrobial ability of the peptide 100µl of supernatant was spotted onto a Muller-Hinton (MH) agar plate that had been inoculated with *E. coli* DH5α (used as a model) and incubated

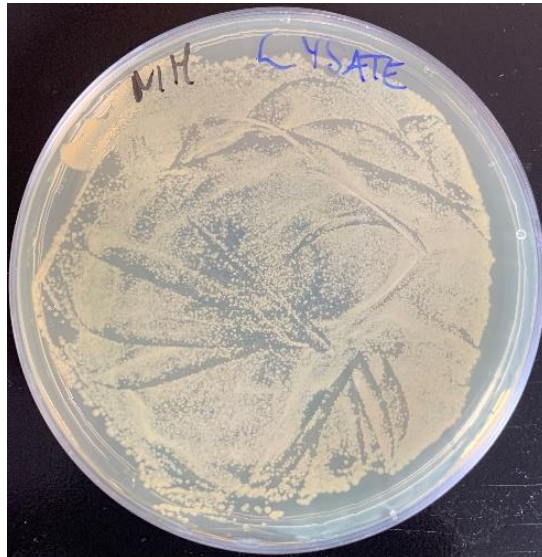


overnight at 37°C. The plate showed no signs of inhibition from the protein in the supernatant. This could be that the fusion protein shows no antimicrobial properties, is not at sufficient concentration, MSBD1 needs to be cleaved from the MBP tag or because of the potential evidence of forming aggregates is not properly folded.



**Figure 7.6 Gene expression and solubility testing of MSBD1**

*12% SDS-PAGE of expression profile and solubility test stained with Coomassie blue stain. Lane 1 Protein Ladder; Lane 2 Expression culture pre induction with 1mM IPTG; Lanes 3-6 post induction profile 1-4 hours; Lane 7 Cell pellet post lysing procedure; Lane 8 Supernatant post lysing procedure.*



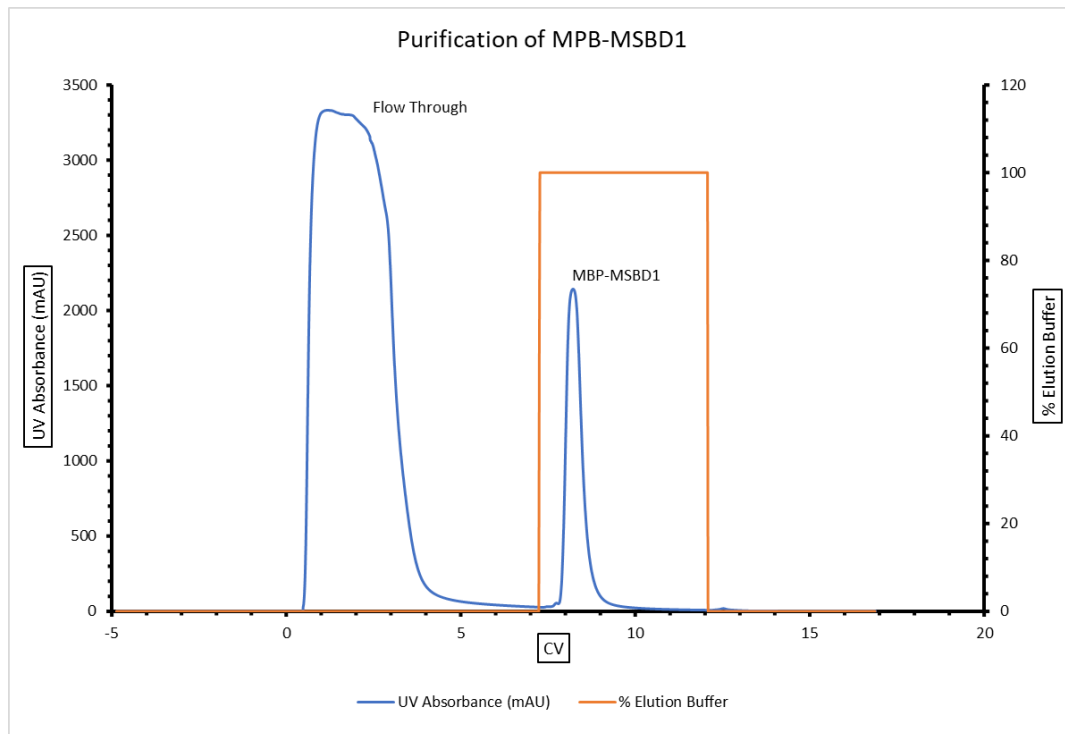
**Figure 7.7 Initial antimicrobial testing of *E. coli* DH1α/MBP-MSBD1 lysate**

*MH* plate with *E. coli* DH5α bacterial lawn showing no sign of growth inhibition. 100μl of lysate was spotted onto the central part of the plate and grown overnight at 37°C.

### **7.5 MBP affinity chromatography purification**

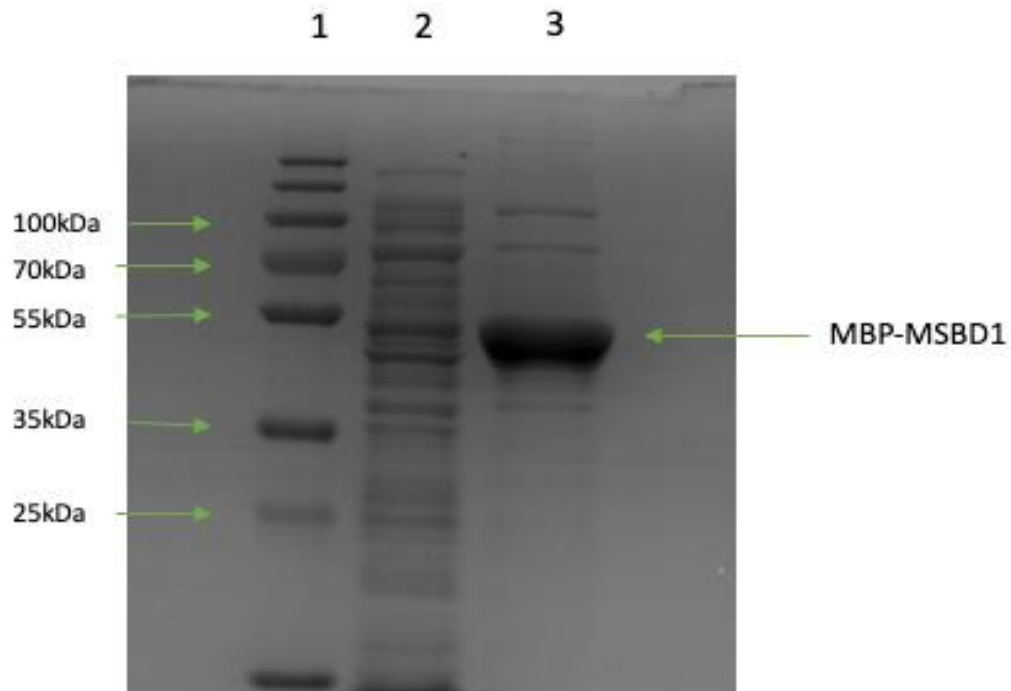
Affinity chromatography using Hi-trap MBP 1ml column was performed on 5ml of supernatant. Affinity chromatography was successful. By loading the clarified lysate onto the column in the absence of maltose the fusion protein bound to the column. This was eluted off the column by isocratic elution with an elution buffer containing 10mM maltose (refer to methods section 9.1.14). This is shown in the chromatogram in figure 7.8. A single peak can be seen that corresponded to the elution of the MBP-MSDB1 fusion protein. The fractions that were collected were 1 ml covering the flowthrough and the elution peak once the UV absorbance had risen over the baseline by 50mAU. The elution fraction was approximately 1.5ml. These fractions were visualised by taking 20μl of each and using a 12% SDS-PAGE gel the second elution peak was confirmed to be the fusion protein. Also, it can be seen is some contamination within the sample and this could have been due to some non-specific binding from unknown protein within the sample. Overall, the purification was sufficient to perform the cleavage of the tag to perform further downstream processes.

As with the supernatant to try and establish a potential screening process 100µl purified fusion MBP-MSBD1 protein from the elution fraction was spotted onto a MH agar plate. This did not show any antimicrobial properties against the *E. coli* DH5α model strain (Figure 7.10).



**Figure 7.8 Affinity chromatography of MBP-MSBD1 UV absorbance trace.**

Chromatogram showing the UV trace (blue) of the purification of 5ml of 0.2µm filtered lysate. 5 CVs of 100% elution buffer (orange) was used as a step gradient to elute the MBP-MSBD1 fusion protein. The fusion protein came off in one sharp peak of approximately 1.5CVs. The flow through fraction between 1-2CVs was used as sample for the SDS-PAGE.



**Figure 7.9 SDS PAGE gel electrophoresis of Affinity Chromatography**

12% gel with fractions collected and stained with coomassie blue. Lane 1 Protein ladder; Lane 2 flow through fraction; Lane 3 Affinity purified MBP-MSBD1 fusion protein.



**Figure 7.10 Antimicrobial testing of purified MBP-MSBD1**

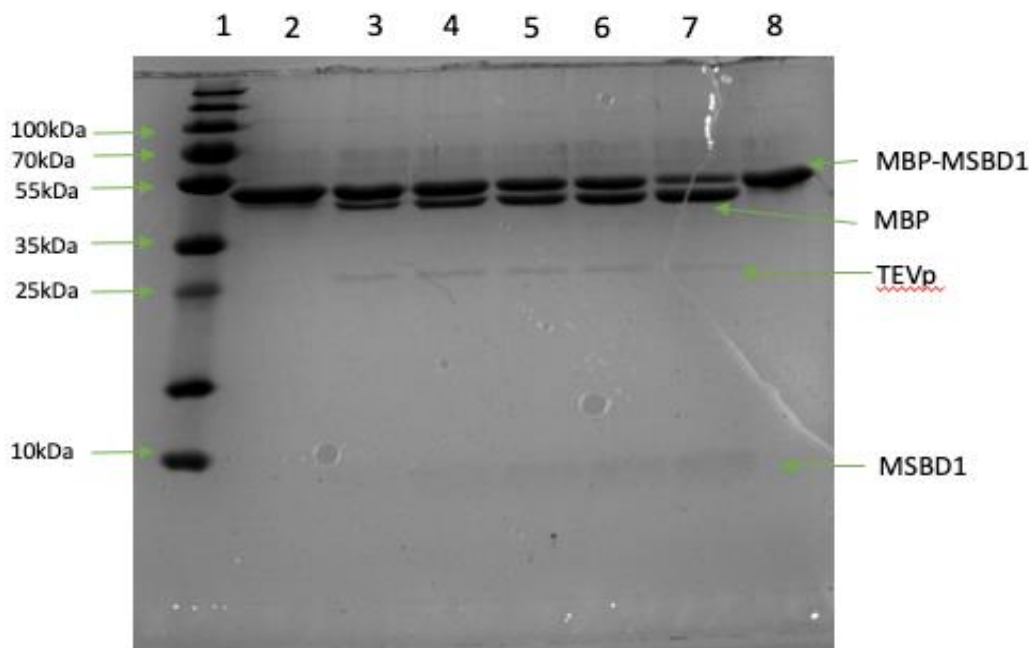
MH agar plate with *E. coli* DH5 $\alpha$  model strain showing no sign of growth inhibition from 100 $\mu$ l of affinity purified MBP-MSBD1 fusion protein. Plate was grown overnight at 37 $^{\circ}$ C

### **7.6 Tobacco Etch Virus Protease (TEVp) cleavage of fusion MBP-MSBD1**

Once the MBP-MSBD fusion protein had been purified using affinity chromatography TEVp was used to cleave the MBP tag from the MSBD1 peptide. 5µl of TEVp (50IU) was able to cleave the peptide from the tag with some degree of efficiency, however this was not achieved to 100% (figure 7.11). After 18h at 30°C there was still some fusion protein remaining. For the MSBD1 protein to be without any overhang from the cloning into the MCS the sequence which was used had a different amino acid in position 1. This is in contrast to the manufacturer's published sequence for the cleavage site, however in a study by Kapust *et.al.* (2002) they showed that the specificity of the amino acid in position 1 was not detrimental to the recognition of the protease to cleave at this site. It was noted that if the amino acids in the other positions (2-6) were altered then this would be detrimental to the efficiency. Another reason could be that the digestion was not performed for a long enough time. In any case, the fusion protein was cleaved to allow further downstream processing.

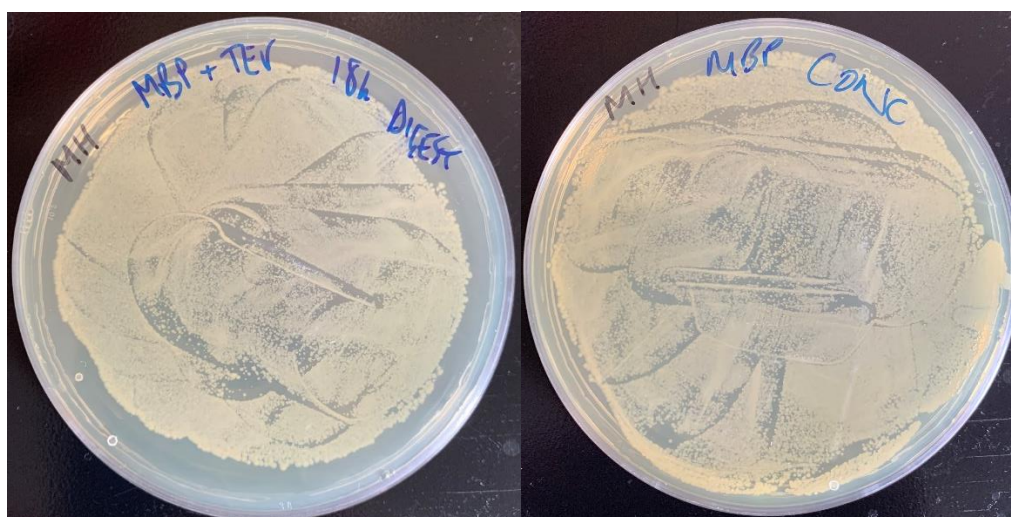
As with the other processes outlined earlier this cleaved mixture of peptides was tested against the model strain to investigate whether the free peptide had any antimicrobial activities. 100µl of this mixture was added to a MH agar plate, however there was no inhibition of growth (figure 7.12).

With the use of Tris-Glycine SDS-PAGE it was very difficult to visualise the low molecular weight of the MSBD1 peptide. It was only 4.5KDa and this was too small for the platform that was used in this study. The move to a more appropriate gel system would be extremely beneficial especially since the downstream processing of the peptide from this point onwards would require visualisation to further investigate. The recommendation to move to either a Bis-Tris or Tricine protein gel system would be of a huge advantage. Along with the change of protein gels a more sensitive staining method would also contribute to the analysis. A stain such as silver staining would contribute to aiding better visualisation.



**Figure 7.11 Cleavage profile of MBP-MSBD1 with TEVp.**

12-20% SDS-PAGE gel. Cleavage time profile of MSBD1 with TEVp. Lane 1 - Protein ladder, Lane 2 and 8 – Pre addition of TEVp, Lanes 3-6 – 1 hour intervals from 1h-4h, Lane 7 - 18 hours. Digestion reaction was performed at 30°C.

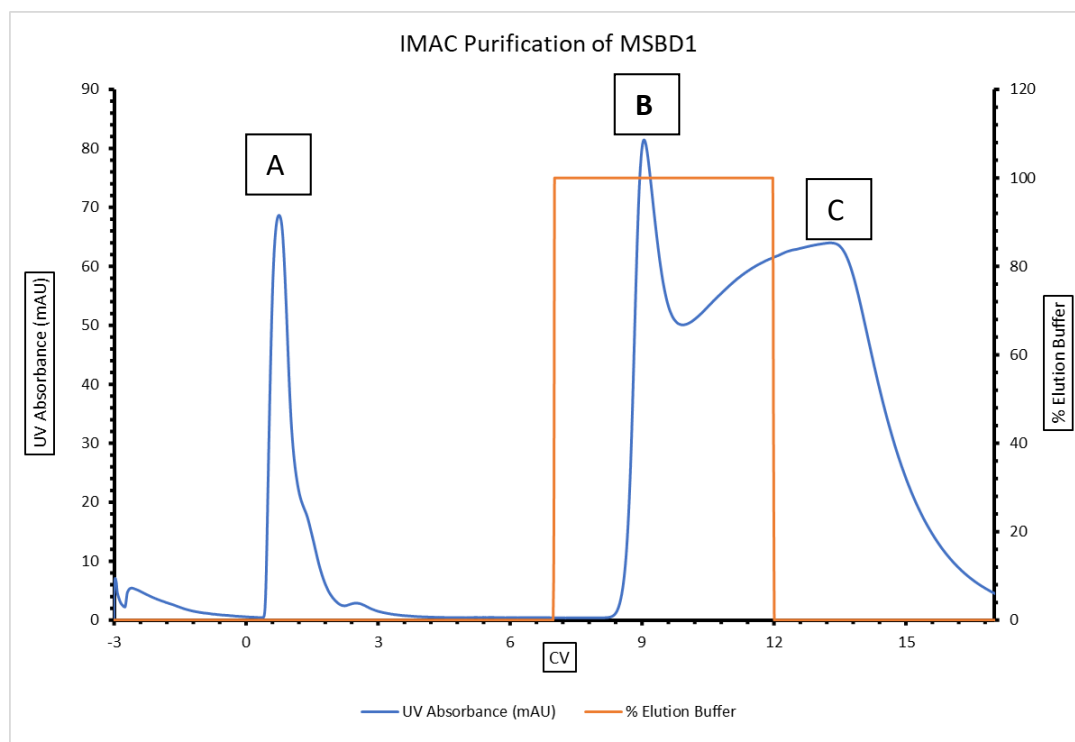


**Figure 7.12 Antimicrobial testing of cleaved MSBD1 mixture.**

MH plates with *E. coli* model strain bacterial lawn showing no signs of inhibition. Plate 1 with 100µl of cleaved MBP-MSBD1 fusion protein mixture and plate 2 showing concentrated/diafiltered mixture. Plates was grown overnight at 37°C.

## 7.7 Purification of MSBD1 with Immobilised Metal Affinity Chromatography (IMAC)

As part of the purification design process the MBP tag and the TEVp that were used had a poly-histidine tag incorporated into their structure to allow an easy and convenient way to purify the MSBD1 peptide through IMAC chromatography using a flow through method which would bind the MBP and TEVp to a nickel ion IMAC column and allow the peptide to be captured in the flowthrough. The chromatogram in figure 7.13 shows the UV trace along with the flowthrough peaks (A) and elution peak (B). Imidazole absorbs UV light and this can be seen in peak C as the elution buffer is coming through the column. These peaks were collected in a fraction collector and SDS-PAGE was used to see confirm what proteins were in the peaks.



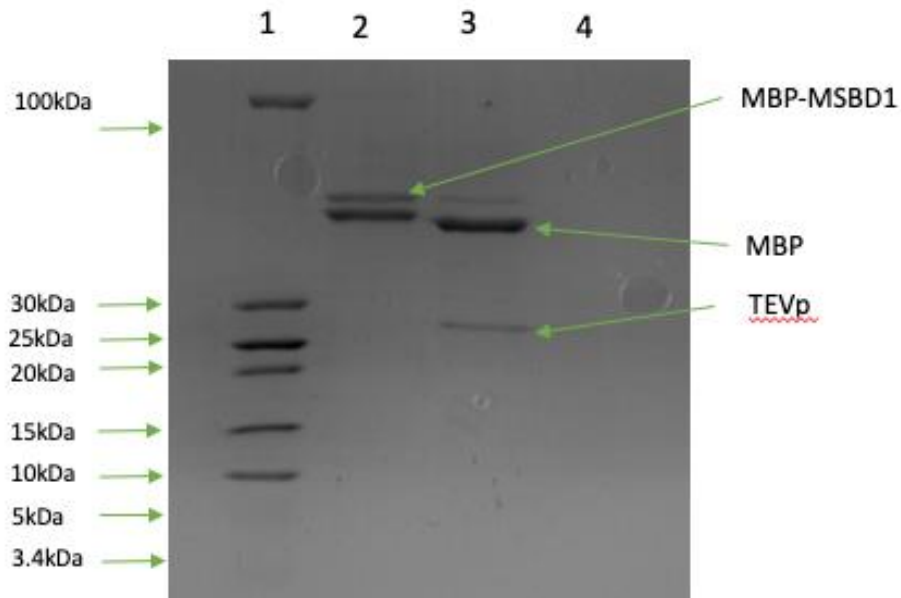
**Figure 7.13 IMAC purification chromatogram.**

Blue line is absorbance curve at 280nm and orange line is gradient showing % of elution buffer. 1 ladder; 2 FT peak A; 3 eluate peak B; 4 peak C Imidazole showing UV absorption and not a protein peak.

The SDS-PAGE gel (figure 7.14) shows several proteins. In peak A, Lane 2 of the gel, two bands corresponding to the fusion protein MBP-MSBD1 and below that the MBP tag. This is abnormal as the poly-histidine tag should allow these proteins to bind to the column. This suggests, as earlier hypothesised that some abnormal folding or aggregation is occurring which is causing steric hindrance of the tag and that it is inaccessible to the sites of the column. Time constraints also dictated that this method of purification could not be optimised.

The gel also showed an absence of the MSBD1 peptide; however, this is not to say that it was not present in the flowthrough, it is probably likely that the Coomassie stain and use of Tris-Glycine gel may have made this visualisation implausible. Even so, because this IMAC chromatography method was designed to allow the MBP tag and TEVp to be bound to the column and MSBD1 to flowthrough, due to this lack of binding by the MBP-MSBD1 fusion and MBP tag, MSBD1 was unable to be purified by this chromatography platform. Peak B from the chromatogram (Lane 3) shows that some of the MBP- MSBD1 fusion protein and MBP tag had been bound by the column and TEVp had been successful in binding allowing TEVp to be fully taken out of the mixture. Lane 4 on the SDS-PAGE was loaded to confirm that peak C was just the absorbance from the imidazole in the elution buffer. Since this method of purification was not ideal, Li and Leong (2011) stated that for successful purification and refolding of the aggregates a refolding size exclusion chromatography procedure would be employed to alleviate the issues outline above.





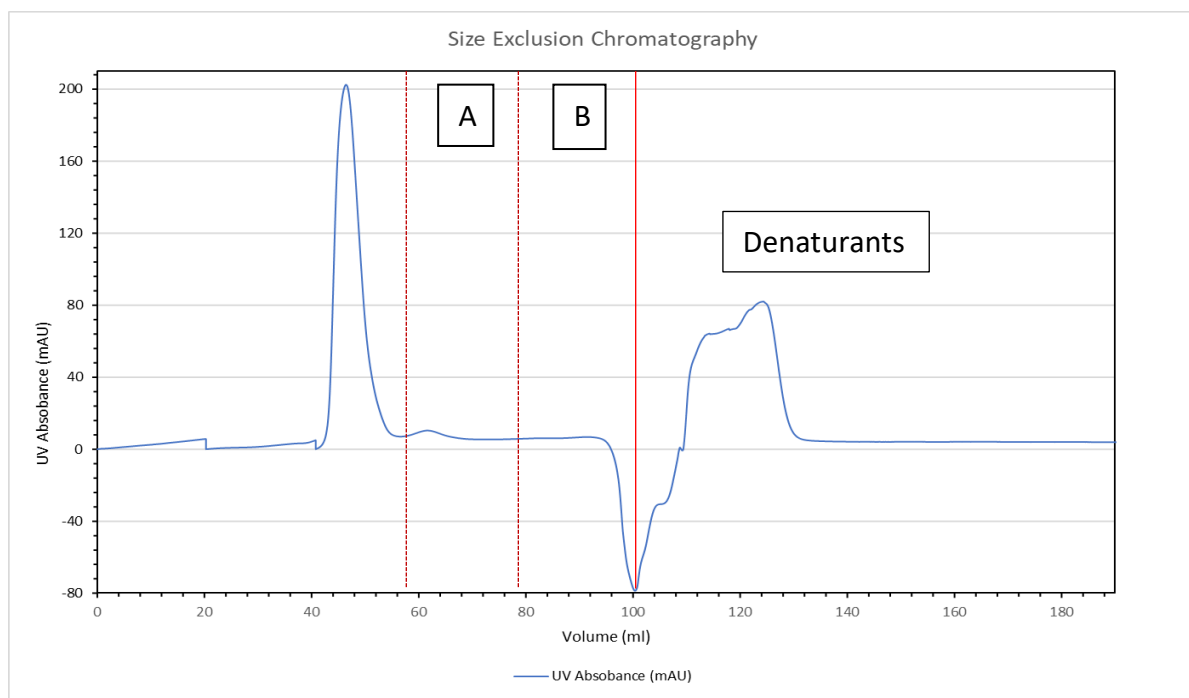
**Figure 7.14 SDS-PAGE of fractions collected during IMAC chromatography.**

Lane 1 ladder; Lane 2 FT peak A; Lane 3 eluate peak B; Lane 4 peak C imidazole showing UV absorption and not a protein peak 12-20% SDS PAGE stained with Coomassie Blue.

### **7.8 Refolding and Purification of MSBD1 using Size Exclusion Chromatography**

It has been reported that high salt concentrations have an inhibitory effect on the peptides antimicrobial ability and that the peptide must be of a certain concentration to influence microbial growth (Crovella, S. *et al.* 2005). The post cleavage mixture was diafiltered and concentrated using a Sartorius Vivaspin 20 to decrease salt concentration and increase the concentration of the peptide. As a qualitative screen to see antimicrobial activity 100µl of this mixture was spotted onto a MH plate to assess its activity. Once again it did not show inhibition of growth (figure 7.17). It was hypothesised that the protein may be misfolded as stated in a previous similar study by (Li and Leong 2011). Following the author's recommendations to perform a refolding of the peptides to produce a correct monomer of the MSBD1 peptide, a refolding of the protein whilst simultaneously purifying the mixture was proposed and executed using a size exclusion chromatography platform adapted from the procedure outlined in (Li and Leong 2011).

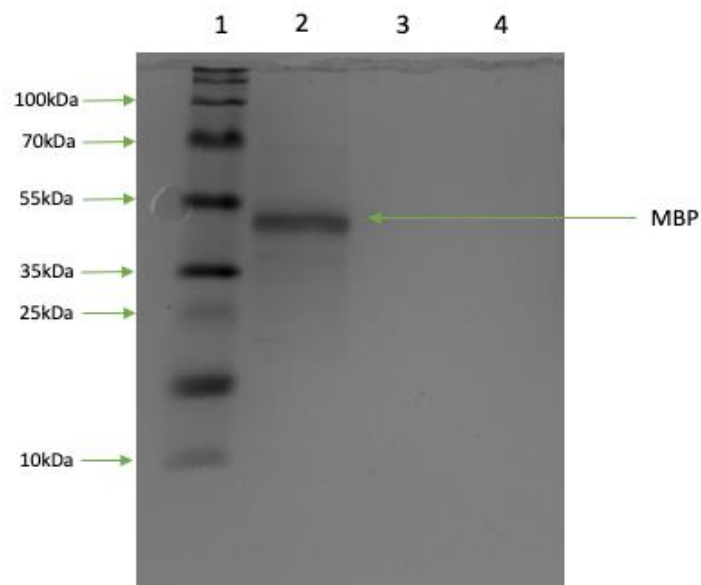
The HiLoad® 16/600 Superdex® 75 pg column used showed separation of the peptides, however one major issue was seen. The refolding buffer had a high base line of absorbance at 280nm and therefore because of the low concentration of the TEVp and MSDB1 within the mixture it was difficult to see the corresponding peaks on the chromatogram (figure 7.15). Fractions A and B were collected were pooled and concentrated and diafiltered to determine where the proteins came through the column (figure 7.15). The first peak in the chromatogram was the cleaved MBP and this is shown in lane 2 on the SDS-PAGE gel (figure 7.16). The pooled fractions A and B shown on the chromatogram were loaded into lane 3 and 4 on the SDS-PAGE gel. These samples, however, did not show up on the stained gel. This could be because the proteins were not high enough concentration to allow sufficient staining or that the VIVA spin with a MWCO of 3000kDa was not small enough and the MSBD1 peptide was lost during the concentration and diafiltration process.



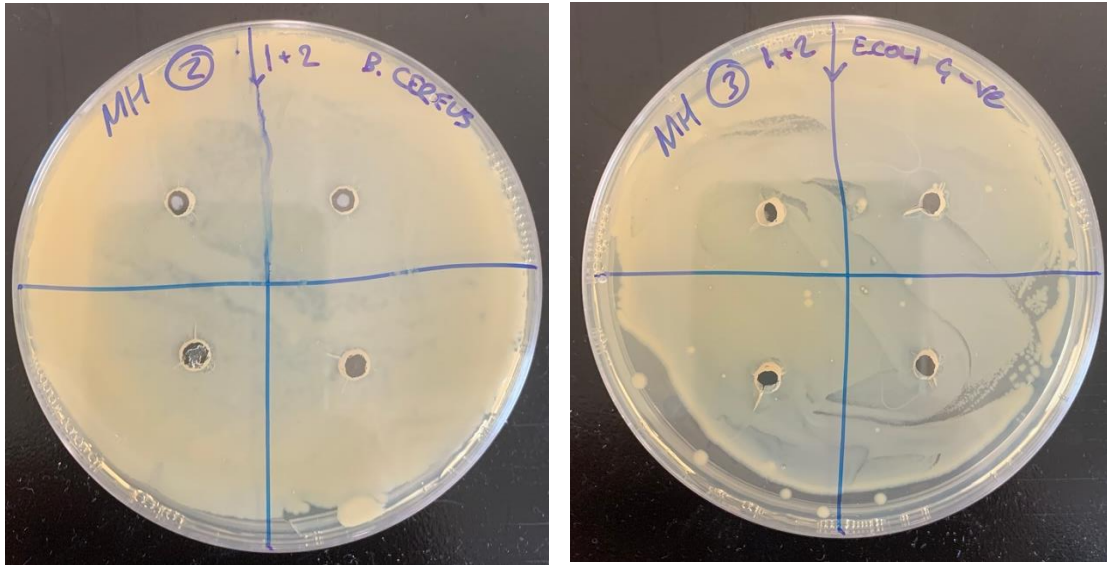
**Figure 7.15 Chromatogram of Refolding Size Exclusion chromatography.**

*The first peak is the cleaved MBP from the fusion MBP-MSBD1 protein. Pooled fractions A and B are shown by red dashed lines. The last dip and peak are the different denaturants coming through the column. UV absorbance at 280nm is blue trace line.*

The pooled fractions that were concentrated and diafiltered were used to qualitatively assess whether the MSBD1 peptide showed any antimicrobial activity after it had been refolded on the column and to also assess, as up until now had only been tested on *E. coli* a gram-negative bacterium, whether it would be active against a gram-positive organism. Therefore, a well diffusion test using *Bacillus cereus* was performed (figure 7.17). 50µl of the concentrated and diafiltered pooled fractions were loaded into wells on an MH plate containing a lawn of *B. cereus* and *E. coli* to test this hypothesis. This showed that there was no activity towards both organisms tested. One question that may arise is that as no absorbance was seen on the trace is the MSBD1 protein present? As maltose binding protein was seen on the chromatogram and the starting material for SEC was a mixture of MBP, TEVp and MSBD1 the protein should have been present in the fractions as these would have eluted much before the denaturants. At this time the study ceased, and no further development could be performed.



**Figure 7.16 20% SDS-PAGE gel showing the fractions from the SEC chromatogram.**  
Lane 1 – Ladder, Lane 2 - Large first peak, Lane 3 – Pooled fraction A, Lane 4 – Pooled fraction B.



**Figure 7.17 Antimicrobial testing of SEC pooled and concentrated fractions.**

MH plates with *B. cereus* and *E. coli* bacterial lawns. The left-hand side of the plates is the concentrated and diafiltered pooled fractions A and the right-hand side is B. The wells had 50 $\mu$ l of the pooled fractions added and were grown overnight at 37°C. Neither plate shows area of inhibition suggesting that that the peptide is not active towards these species.

### 7.9 Summary

The successful DNA extraction procedure, followed by the isolation of the MSBD1 gene of interest from the genomic DNA using the hybrid primer pair designed for restriction free cloning into the plasmid was achieved. With this isolated gene fragment with flanking plasmid annealing sequences, cloning into the plasmid was fast and efficient with the whole process taking a few hours without the need for restriction enzymes further adding a cost saving. The method of restriction free cloning allowed the gene of interest to be inserted into the plasmid next to the TEVp cleavage site without any extra bases in the sequence subsequently adding amino acids into the peptide as would happen with the use of restriction enzymes. One issue to keep in mind when using this method is to allow sufficient time when adding the Dpn1 restriction enzyme to fully digest the parental plasmid as this would cause problems downstream. The protocol of electroporation used in this study also was streamlined giving a total time of two days to transform the plasmid into the expression strain. This required minimal reagents again giving a cost efficiency to the process.

Using the expression system, designing an MBP-MSBD1 fusion protein that could be later cleaved to allow purification of the protein had both advantages and limitations. One such advantage would be that the MBP has high solubility and when overexpressing proteins on bacterial expression system can cause the protein to form insoluble aggregates or inclusion bodies that are difficult to resolubilise and purify. MBP is also easily purified out of the supernatant using Affinity Chromatography with high specificity using an amylose resin and eluting using maltose in the buffer system. By introducing a TEVp cleavage site into the protein sequence allowing removal off this affinity tag also has a distinct advantage in being able to liberate the MSBD1 from the fusion protein. The overall performance of purifying and cleavage using TEVp was suitable, proving proof of concept up until this point in the process.

The manufacturer integrated a poly-histidine tag on both the TEVp and the sequence of the MBP fusion protein plasmid construct to allow efficient purification of MSBD1 by flow through in tandem with binding the TEVp and cleaved MBP to an IMAC column. The results of this need more optimisation as there was residual uncleaved MBP-MSBD1 and another contaminant in the flowthrough. This could be because the buffer conditions weren't optimal or that the fusion protein was causing aggregation and therefore steric hindrance obstruction the poly-histidine tag being accessible to the binding site on the column.

It was noticed in a previous study (Li and Leong 2011) that even though MBP produced soluble proteins they may be misfolded and cause aggregation. After testing its potential anti-microbial properties and showing negative activity against *E. coli* it was hypothesised that this may be a cause and a refolding/purification, and the steric hindrance observed in the IMAC purification strategy was devised. Using similar parameters outlined in the Li and Leong study.

It was difficult to full assess this part of the downstream process due to several issues in analysing the different steps of the process. Firstly, the background base line on the SEC chromatogram was higher than the absorbance that the cleaved MSBD1 and TEVp produced during the run. This was likely because of the low concentrations of the peptides. This made it difficult to see what fraction of the run these came through the column. The fractions were pooled and concentrated to see if these were visible on an SDS-PAGE gel. Again, these were not visualised. These pooled fractions were utilised for assessing if there was any antimicrobial activity, and they did not show any positive results.



## Chapter 8 - Discussion

The choice to study the beta-defensin clusters in reptiles in this work represents a gap in the knowledge we have in this area of research with only a few examples of work in the literature (Benato *et. al* 2013; Dalla Valle *et. al* 2012; Van Hoek *et.al* 2019; Santana *et.al* 2021). Reptiles offer a unique insight into the beta-defensin repertoire and with more knowledge applied to the current understanding may offer insights into the evolution and molecular properties of this kingdom of creatures. Reptiles being the third largest group of vertebrates offer a link between fish and amphibians and birds and mammals, thus they are an interesting model to investigate their immune functions. Reptiles inhabit many different ecological niches and with this provides an interesting vision into their immunological assets. For example, the *Anolis carolinensis* can shed its tail in self-defence and whilst suffering a huge physiological trauma, can resist infection of the wound, allowing healing and regeneration of their tails (Alibardi 2010). Another example is that the Komodo Dragon has a high pathogenic load within its saliva for which it has used to its advantage when hunting prey without being overwhelmed by these microbes, providing a terminal bite for which their prey succumbs to sepsis. These examples show that the immune systems of reptiles are robust and warrant further investigation.

Genomic sequencing has come on leaps and bounds in the advent of next generation sequencing (NGS). NGS has provided an ever growing number of high quality, publicly available genome assemblies. With this, gains in genomic analysis have materialised and have provided the raw data within this work. The bioinformatic methodology employed in this work to search the available genomes for novel Beta-defensin cluster regions and to annotate the genes within these regions had several advantages and disadvantages. By their very nature beta-defensins are difficult to search for using traditional methods such as the BLAST tool because the size of the exons are particularly small, however this was used for initial searches. This technique was enhanced using a concatemer approach whereby strings of known beta-defensin amino acid sequences were employed as a search query. This approach increased the likelihood of finding a match. A clear increase of matches was produced when compared to using a single known beta-defensin when using as a query in a BLAST search. With the publication of literature by van Hoek *et al.* (2019), information regarding the flanking

genes came to light. In this paper they suggested the CTSB and TRAM2 or XPO1 genes flanked the gene cluster in the Komodo Dragon and armed with this information the process to search a genome started with finding these two genes. Then the regions could be further probed using the concatemer approach for the beta-defensins that resided between these flanking genes. This improved the speed at which the gene cluster regions could be finalised. Once the region had been identified the raw sequence data was downloaded, the repeat sequences masked using the RepeatMasker program and the RepeatMasker output translated into a 6-frame output which permitted the sequences to be analysed manually at a DNA coding level. Every match identified from the BLAST search was manually annotated and highlighted on this output for downstream annotations through additional tools. There was a disadvantage to this when it came to finding the first exons by the BLAST tool approach therefore gene finding programs were employed to identify these exons. On top of this, as the first exons were somewhat conserved, the already identified first exon DNA sequences could also be inputted as a query in iterative BLAST searches. Finally, the output of RepeatMasker of identified repetitive sequences within the cluster region was utilised to search gaps that were over 3000bp between repeats for Beta-defensin sequences. The major drawback of this method of annotation is that it was extremely time-consuming, however, a large advantage is the confidence you gain knowing that such an in-depth, sequence level analysis ensures that all unknown beta-defensins were identified. Splice sites were identified by two different splice site programs, which gave an extra level of confidence on exon boundaries. The two programs use different approaches and if a consensus was found then this gave reassurance to the finalised beta-defensin sequence. Santana *et al.* (2021) employed a more automated approach to searching for Beta-defensin cluster in crocodylian genomes. Their process made the use of Hidden Markov Modelling (HMM) where a consensus sequence is produced by combining known beta-defensin sequences and subjected the genome to this. One drawback of this is that it is only powerful if you have many sequences to produce a strong robust consensus. When comparing their sequences with the ones identified in this work a few differences were identified. With the lack of manual DNA analysis at sequence level a few anomalous results were presented such as genes which weren't present within the cluster region, had extra exons which were not present in the genomic sequence, and genes identified which did not have an initiating methionine. Therefore, it could be argued that the



disadvantages of a manual, time consuming approach to identifying beta-defensins would outweigh the advantages of a quick automated process.

Genomic location, organisation, gene number and variation along with some downstream analyses have provided an expanding picture into the beta-defensin cluster regions of reptiles. The genome assemblies in this work were chosen because they were at chromosomal assembly level or at a scaffold level so that it provided the data to completely construct and annotate the beta-defensin cluster region and to provide information into the location and the genes that flank these regions. In the genomes of the lizards *P. muralis* and *L. agilis*, the turtle *C. mydas* and tortoise *G. evgoodei*, where chromosomal numbers had been confirmed and assembled, the cluster regions resided on chromosome 3. This is the same as the chicken (Hellgren and Ekblom 2010). However, in the snakes, *C. v. viridis* and *N. naja* they cluster region was identified on chromosome 1 and *T. elegans* on chromosome 4. One could hypothesize that these clusters translocated to these chromosomes when snakes diverged from the lizard's lineage. The same could also be hypothesised for *T. elegans* separation of the *Colobridae* from the *Elepidae* clade. This, however, with further analysis on more species would build a truer to life picture of the chromosomal location of the beta-defensin cluster. Reptilian beta-defensin genes have been shown to reside in a single gene cluster (van Hoek *et al.* 2019; Santana *et al.* 2021) which is analogous to birds (Chen *et al.* 2015, Hellgren *et al.* 2010, Cheng *et al.* 2015) and this is shown in the beta-defensin clusters identified here. Research conducted by Cheng (2015) showed conserved synteny between beta-defensin clusters and demonstrated that Cathepsin B (CTSB) and translocation associated membrane protein 2 (TRAM2) flanked all the defensin regions in birds. In testudines and crocodylians the beta-defensin cluster has been shown to share this synteny, which was confirmed in this analysis. However, it was, until this work, unknown if TRAM2 was the flanking gene in lizards and snakes (van Hoek *et al.* 2019). This analysis proved that the flanking gene in squamates confirms that the synteny is not conserved for this order of reptile and that Exportin 1 (XPO1) gene flanks the cluster region. This finding will allow easier identification of novel defensin clusters in squamates.

The number of beta-defensin genes that have been identified in vertebrates is exceptionally variable. Tu *et al.* (2015) performed *In silico* analysis of 29 genomes and it was predicted that the number of whole genes ranged from 1 in the western clawed frog to 20 in cattle. 14 have been identified in birds with exceptions like the Zebra finch having 22 Beta-defensin genes

(Hellgren *et al.* 2010). This is also the case in reptiles; 37 have been identified in the green anole lizard (Dalla Valle *et al.* 2012), 26 identified in the Chinese soft-shelled tortoise (Yu *et al.* 2016) and 66 variants identified in the Komodo dragon (van Hoek *et al.* 2019). This study has uncovered similar results. Lizards showed the greatest number of beta-defensin genes ranging from 80 potentially novel defensins in *P. muralis* to 64 in *L. agilis* and 34 in *Z. vivipara*. In snakes the number was also highly variable; 15 were identified in *C. v. viridis*, 27 in *N. naja* and 51 in the *T. elegans* cluster. Within the testudines *G. evgoodei* had 47 novel genes and *C. mydas* had 39. One group identified 18 new beta-defensins in *A. mississippiensis* and 22 from *C. porosus* (Santana *et al.* 2021). In their work they identified these through hidden Markov modelling (HMM), however, the method employed in this work contradicted their predications as only 14 potential genes were identified in *C. porosus* residing between CTSSB and TRAM2. The genes identified in this study also display slight differences in sequence compared to Santana *et al.* which could be a limitation of the HMM method as more in-depth manual sequence analysis is required to fully characterise the small sequences at a DNA level. The structure of beta-defensins is usually described as being ribosomally synthesised as a preprodefensin. This represents a signal peptide that is usually a leucine-rich, hydrophobic alpha-helix (pre-peptide) and intermediate region denoted the pro-domain and finally a mature active peptide, responsible for the peptide ability to show antimicrobial properties, with a common three beta-sheet arrangement involving the common cysteine defensin motif. These then undergo post-translational modifications such as cleavage of the different parts of the peptide to generate the final mature active antimicrobial peptide. In most cases the signal peptide, along with the pro-domain is encoded by the first exon and the mature active peptide is encoded by the second exon (Ganz 2003). This work shows that the genes uncovered seem to follow the general rule for the organisational structure of beta-defensins. Lizards follow a similar pattern to that described by Dalla Valle *et al.* (2012) on the work that was conducted on the green anole lizard where there was evidence of a three exon structure. *L. agilis* and *Z. vivipara* exhibited 3 potential beta-defensins with this structure, however it must be noted that because the coding sequences for the third exon are typically only a few amino acids long the sequences are hard to detect *In silico* and therefore this is a limitation of this workflow. In snakes the signal peptide is encoded by the first exon and the mature peptide is encoded by the second exon, however, they seem to lack the presence of a large pro-domain or may even lack this domain completely. In the testudines they seem to follow

the two exon structure which was also described in the Chinese Soft-shelled Tortoise (Benato *et al.* 2013). Crocodylians also have a two exon structure too, however, Santana *et al.* (2021) describes a four exon structure with the first and second exon encoding the 5' and 3'-UTR respectively. With the character of beta-defensin genes having relatively small exons this provides challenges when identifying these and their boundaries *In Silico*. Nevertheless, definitive answers will come from transcriptome analyses when the data becomes available. All the defensins uncovered have the typical beta-defensin conserved cysteine motif in the mature peptide which allows for the intramolecular disulphide bonds to form. Within this sequence a Glycine-X-Cystine arrangement is somewhat conserved in this sequence. This arrangement is thought to be responsible for forming a 'bulge' that allows for the correct folding and native structure within the mature peptide (Tu *et al.* 2015). Additionally, a conserved signal peptide was observed in all the beta-defensins. Most of the beta-defensins had a small or non-existent pro-domain, however, in lizards, testudines and crocodylians exhibited a few defensins within their cluster regions that possessed large pro-domains in the primary sequences. These were encoded by the first exon in all cases. Recently long pro-domains were described in crocodylians (Santana *et al.* 2021; Tang *et al.* 2018). Tang *et al.* analysed the beta-defensins with long pro-domains from the Chinese alligator and found them to be very similar to mammalian alpha-defensins in their physical properties, sharing comparable net charges, hydrophobicity, and length. This observation could suggest there could potentially be a unique connection between these long pro-domain beta-defensins and their alpha defensin counterparts.

It was shown that in alpha-defensins this long pro-domain balances the charge of the mature active peptide potentially keeping it inactive during synthesis and reducing host cytotoxicity (Michealson *et al.* 1992). In the beta-defensins possessing this long pro-domain region was investigated and it was found that this charge balance was broadly similar to alpha-defensins identified in the studies outlined above. As mentioned earlier beta-defensins undergo several post-translational modifications. Ganz (2003) describes how the different domains of the prepropeptide of alpha-defensins in myeloid cells undergo changes as the peptide transitions from being newly synthesised through to becoming a mature active peptide. Ganz describes that during defensin synthesis the signal peptide is rapidly removed with subsequent proteolytic processing with final cleavage of the pro-domain into the mature peptide occurring in the maturing granules. Once matured the neutrophils are released into

the blood whereby these granules with high defensin concentration assist in degradation of microorganisms during phagocytosis. This sequential cell sorting may provide the protection against cytotoxicity, hence the evolutionary result of this long pro-defensin region. It has been found that these long pro-domain beta-defensins are highly expressed in the organs of the digestive tract (Tang *et al.* 2018) with a similar expression pattern observed in mammalian alpha-defensins (Selsted and Ouellette 2005) and therefore could offer insights into the evolutionary link between the reptilian long pro-domain beta-defensins and the mammalian alpha-defensins. More characterisation studies would further elucidate these similarities. Mature active beta-defensins are often described as being cationic, however, several of the reptilian mature beta-defensins found in this work possess an overall net negative charge. All the species analysed in this work were found to hold these beta-defensins in their cluster regions with the lizards having the most numerous anionic mature peptides. This has been described in the Chinese alligator (Tang *et al.* 2018), Komodo dragon (van Hoek *et al.* 2019), Green Anole lizard (Dalla Valle *et al.* 2012). This could be the first time that anionic defensins have been identified in snakes and testudines.

Observations from the selection analyses conducted in this work suggests that reptilian beta-defensins have high rates of gene duplication events with high rates of sequence variability which results in numerous orthologues and paralogues among the different species to give rise to a large difference in total gene number. Several mechanisms may be involved in these differences. The beta-defensin clusters characterised in this study showed varying degree of synteny between orthologous genes within the species groups. The genes closest the CTSS showed the most homology between species but as the genes moved further along the cluster towards TRAM2/XPO1, they became more variable. Within species groups many one-to-one orthologous gene were identified such as what was described in crocodylia (Santana *et al.* 2021) and in the testudines. However, we observe that when looking at orthology in snakes and lizards this becomes less certain and could be down the different rates of duplication/extinction rates that have been previously seen among different species (Semple *et al.* 2005). These differences may prove difficult to describe one-to-one orthology between different orders of reptiles, although more phylogenetic analyses could uncover some of these answers.

It is common thought that gene clusters arise from duplication events such as unequal-crossover and mismatch pairing during meiosis, but this model of gene duplication only

describes concerted evolution and as such doesn't describe the formation of pseudogenisation and neofunctionalisation within these gene clusters. A 'birth and death' model has been used to describe duplication/extinction events within multigene families of the immune system (Nei and Hughes 1992; Nei *et al.*1997). This model describes that the gene duplication and subsequent diversification and pseudogenisation are routine especially within immune system gene families. New genes are created by duplication events and some genes are maintained in the genome for a long period of time whereas some genes are deleted or become non-functional over time to become pseudogenes. Looking at phylogeny of the sequences within the gene cluster evidence for this model is present. This is more noticeable in species where gene copy number is high such as the lizards, snakes and testudines analysed here. The tortoise, *G. evgoodei*, demonstrates this particularly well. The dot plot shows two areas within the cluster with high level of duplications, and this corresponds to the phylogeny where the paralogous genes are grouped into two distinct branches. There are also examples of this in the *P. muralis* cluster region too.

The cluster regions analysed in this work show several different duplication events. Tandem gene duplications have been seen in most of the species but there is also evidence for cluster duplications in some species. In the *T. elegans* cluster region dot plot and within the sequences shown in the multiple sequence alignment corresponds to a region of around 300 kb that has duplicated to give rise to a region containing up to 10 nearly identical beta-defensins in two different tandem groups. Genomic Inversion events have also been identified as another mechanism to generate variation within the cluster regions. This was identified when analysing conservation of synteny in the three lizard species. Each of the cluster regions showed inversion of one or more large segments when compared to one or the other species within this analysis. It seems that within the cluster regions that there are many different mechanisms that play a role in driving variation which would reinforce functional diversification within the beta-defensin genes within these clusters subsequently ensuring the survival of the species when confronted with pathogenic challenges.

In evolution, there is often a back-and-forth direction of change between positive and negative Darwinian selection as different forces apply pressures to the species. Gene duplication is therefore a fundamental process by which novel proteins with novel functions evolve and it is therefore useful to identify which direction genes, parts of genes and individual residues are travelling in, for example how genes may acquire novel biological

functions. This will inevitably help scientists to understand why the genes may have formed this way and help potentially bring the knowledge forward for harnessing these properties into novel uses in real-world situations which may help mankind. To shed light on these processes different analyses were performed to determine what selection pressures are occurring, firstly on different parts of the gene and then looking in finer detail what selection pressure the residues at a sequence level are under. Looking into synonymous (*dS*) and nonsynonymous (*dN*) nucleotide substitutions gives an insight into the evolutionary divergence of the mutations within these duplications (Hughes 1999). Synonymous or silent mutations are usually invisible to natural selection as these are mutations that don't alter the amino acid in the protein sequence as the codons are shared (Akashi 1995), whereas, nonsynonymous mutations change the amino acid codon, which may be under greater selection pressure. In this study we used a statistical method devised by Nei and Gojobori (1986). The ratios were computed using the SNAP program on the HIV database online server which calculated pairwise comparisons of each gene within the cluster. The ratios were plotted to provide a visual representation of the distribution of these ratios. From the signal peptides produced in the IP server and the mature peptide these followed observations in that the signal peptide was undergoing negative/purifying selection and the mature peptide was still under slight neutral selection. The conservation of sequences in the signal peptide are to be expected as this region of the beta-defensin is largely undergoing purifying selection as there is a high degree of homology within these sequences. The interesting observation is that one may expect to see more of a positive selection distribution in the mature peptide as this region of the beta-defensins, involved in host defence, required to change due to the pressure given by host/pathogens dynamics. However, this may be saturated by the cystine motif that is a conserved characteristic of beta-defensins. There may be other pitfalls with this method when looking at paralogous sequences within species. In the literature similar analysis has been reported with more phylogenically robust orthologues along newly duplicated gene clusters within the chicken and zebra finch (Hellgren *et al.* 2010). In addition, an iteration of the test YN00 test, from the PAML program (Yang 2007) involving more recent statistical tests would reveal that some pairwise comparisons were too divergent to bring back a meaningful answer, therefore more development of statistical workflows needs to be done in this area if cluster regions genes are to be analysed with confidence. However, it still provides a useful overall picture of the different regions of the beta-defensins gene selection

pressures. Thus, with these limitations, a different strategy may need to be employed for future work.

Identifying residues within a protein that might be undergoing positive or negative selection might give insights into which residues are important in the biological function. Previous work has shown that positively selected sites are largely restricted to this region with the signal and pro-domain being largely unaffected (Maxwell *et al.* 2003; Morrison *et al.* 2003) so the decision was made to concentrate site wise analysis on the mature peptide. There was considerable variation with the sites undergoing different selection pressures. In Lizards there were between 2-5 sites that were undergoing positive selection and many more, between 17-21 sites undergoing purifying selection. In snakes between 4 and 17 sites within the *T. elegans* second exon are undergoing positive selection and 7-13 sites undergoing negative selection. In testudines 13-18 site undergo positive and between 8-9 site undergo negative selection. Finally, in the crocodilian 2-4 sites undergoing positive selection and 12-13 sites undergoing negative selection. The 17 sites within the *T. elegans* and 17-21 site within the testudines second exons shows that these regions are undergoing significant positive selection suggesting that the beta-defensins may be acquiring novel biological functions which may not be completely understood and fit with the current theories of the role of the beta-defensin active mature peptide. The sites subjected to negative selection were the GxC residues part of the 6 cystine conserved Beta-defensin motif and these tend to be situated between the motif where many of the positive sites were positioned. This could hint that these regions in the 'bulge' confer that they are likely to play a role in its function or shape. Similar findings were found within mouse and human defensins whereby the beta-sheet arrangement with cystine pairing largely unaffected and the loop or bulge regions showing most positive selection (Semple *et al.* 2005). Cheng *et al.* 2015 describes that they detected negative selection (overall mean  $dN < dS$ ) in most of the genes examined where the conservation of amino acid residues was a prerequisite to maintain the functionality of beta-defensins, however positive selection was involved the diversification of these genes at specific codon sites. Duplicated sites are subjected to a greater selection pressure because unnecessary identical copies won't be preserved in the genome, therefore the need to diversify will allow for this retention. On the other hand, single copy genes with many nonsynonymous substitutions in pathogen-binding will ultimately be fatal and as a result some degree of purifying selection needs to be maintained (Chen *et al.* 2015). The

variation of such regions and sites within these peptides also shows that the 'arms race' between pathogen and host is an ongoing battle for survival. Additionally with these findings, novel synthetic defensin-like peptides could be explored for novel antimicrobial functions and activities. Directed evolution techniques along with high throughput analyses could play an important role in the development of new antimicrobials.

Transposable elements and repetitive sequences are a large proportion of the genome (Biémont & Vieira 2006) and play a role in the evolution of species (Cordaux & Batzer 2009). Analysis into the repeat sequences landscape of the Beta-defensin gene cluster regions were undertaken. RepeatMasker with RepBase tetrapod database was used against these regions to build a picture of the different types of repeat sequence and to give a picture on the abundance of the different types within. Between 17-32% in squamates, 49-53% in testudines and 37-39% of the cluster region was highlighted as containing a repetitive sequence within the cluster region. The predominant repeat elements in all the species analysed was Long Interspersed Nuclear Elements (LINEs) and CR1 repeats being the most abundant within this class of retroelements accounting for around 74-79% in testudines, 83-88% in crocodylia and 44-75% in squamates. DNA transposons were also a large proportion of the repeat sequences identified. With these repeat sequences in such abundance in these regions it could be hypothesised that these could be drivers for the evolution of beta-defensin by way of duplication by retrotransposition or unequal cross-over (Zhang 2003). Chen *et al.* (2015) investigated the repeat landscape of the golden pheasant and hwamei Beta-defensin genes and proposed a model of duplication dependant strand annealing model for a gene duplication mechanism in which after a double strand break the broken ends of the DNA strand 'searches' for nearby homologous sequences which are then repaired to for newly duplicated regions within a closely situated DNA region. They propose this mechanism almost down to the ubiquity of the repeat elements within the genomic sequences. Further analysis into this could provide a deeper understanding of how these regions become highly duplicated. It would also be beneficial to see if these regions show a higher proportion of repetitive sequences compared to a genomic level to also add substance to the vital question. Unfortunately, this was not performed as more computing power was required to achieve this.

Conservation of synteny studies between species have been investigated in several studies (Schutte *et al.* 2002; Radhakrishnan *et al.* 2007; Cheng *et al.* 2015; Santana 2021) with only



Santana *et al.* exploring data from reptilian species, namely four crocodylian species, hence there is a need to further explore this within reptiles. This work studied three different reptile clades, three snake, three lizard and two testudine cluster regions for syntenic similarities and differences. Within each cluster there was varying degrees of homology within the gene sequences and genomic organisation. As outlined earlier, the lizards showed conservation of synteny with the flanking genes however there were differences within the cluster region. The genes leading away from CTSB showed a high degree of one-to-one homology but as the gene become more distant from the CTSB this is where more of the variation began to happen. This could suggest that genes were present before the separation into the species that we see currently. Genomic inversions were also present when comparing the three species indicating an unidentified mechanism present for genomic variation to occur. However, these differences add to the evolutionary picture of the defensin region in lizards. The *P. muralis* cluster region also contained several paralogous genes in two highly duplicated regions when compared to the other lizard species and this region could be a useful evolutionary model to explore, why it may have arisen and give insights into the mechanisms of duplication within these clusters in reptiles. The conservation of synteny in snakes also provided an interesting illustration into the variability between species. As before CTSB and XPO1 flanked the cluster regions and as with the lizards the first few beta-defensins leading away from CTSB showed the most homology within the three species. There was a region within the *T. elegans* cluster that showed genus specific beta-defensins which were absent from the other two species. When looking into the phylogeny of the snakes this set of beta-defensins could be isolated to the Colubridae family of snakes and could have arisen with the separation of Colubridae from Elapidae families. In addition, there was a set of beta-defensins that were specific to the *T. elegans* and *N. naja* which could be specific to Colubridae and Elapidae families. With more sequencing of snake genomes in the future this data could be further reinforced. The testudines cluster regions showed a large amount of homology covering the whole cluster region with just one expanded region of paralogues in the *G. evgoodei* cluster. Tang *et al.* (2018) identified a similar pattern in the Chinese Alligator whereby a number of beta-defensin paralogues with long pro-domain regions showed to have predominant expression in the intestinal tract. These are like the identified paralogues in *G. evgoodei* and could serve a similar function.

A method for cloning, expression, and purification of a mature beta-defensin peptide was explored. To allow successful expression of the MSBD1 mature peptide a suitable expression vector system had to be constructed. A maltose binding protein fusion peptide was used as it increases the solubility of the fusion partner and allows easy purification via an amylose resin (Li & Leong 2011; Li *et al.* 2014; Vu *et al.* 2014) and to alleviate the possibility of toxicity to the expression strain. The peptide was tagged to MBP with an intermediate Tobacco Etch Virus protease (TEVp) recognition sequence to liberate the fused MSBD1 peptide. A restriction free cloning method was used to allow precise insertion of the sequences into the vector which allowed the peptide to have the exact sequence as the native sequence from the isolated gene from the genomic sequence. This method worked very well, expression was achieved in the soluble fraction and purification through the amylose resin was accomplished with high purity. The TEVp was able to isolate the peptide with efficiency. The Shuffle® (New England Biolabs) *E. coli* expression strain was used for its efficiency to produce proteins with the correct folding and cysteine pairings. Despite being expressed there were no in-house apparatus which would allow the correct cystine pairing to be confirmed. At this point in the process, optimisations would be needed to produce enough protein for further analyses. To successfully visualise the small 4.5kDa MSBD1 peptide through conventional SDS-PAGE with Coomassie staining a Tricine gel with silver staining would be needed to confirm the presence of the peptide. Another limitation in the characterisation of the peptide was to establish if there were soluble aggregates. Li & Leong (2011) described their peptide, hBD25, using size exclusion chromatography to deduce the size of their peptide in its native states. This could be why MSBD1 did not show any antimicrobial properties when it was cleaved from the MBP fusion partner. Possible supplementary evidence for this is that the MBP and TEVp have a poly histidine tag which can be separated from the cleaved MSBD1 peptide by immobilised metal affinity chromatography (IMAC). When this was performed on the post cleavage mixture the MSBD1 peptide should come with the flowthrough and the MBP and TEVp should bind to the column, thus allowing purification of the MSBD1 peptide. However, there was evidence of uncleaved MBP-MSBD1 fusion protein in the flowthrough. This could be due to steric hindrance of the poly histidine tag by the formation of the soluble aggregate. Therefore, a refolding strategy was performed on the post cleaved mixture of peptides. Once the mixture had been denatured the refolding of the proteins on size exclusion column was performed. There was an issue of the absorbance of the low concentrations of TEVp and

MSBD1. As the refolding buffer had a raised background absorbance there was not sufficient concentrations of peptide in the sample to rise above this and see a definitive peak on the chromatogram. Along with not being able to visualise the peptide through SDS-PAGE this made deducing which fraction contained the peptide very difficult. Therefore, a pooling of fractions with concentration and diafiltration was undertaken. This had two functions. Firstly, to identify if the MSBD peptide was isolated and if there was sufficient concentration of peptide to show on SDS-PAGE to allow antimicrobial screening tests. Secondly to buffer exchange into a lower salt buffer as it has been shown that high concentrations of salt affect impact the efficacy of the antimicrobial ability of beta-defensins (Crovella *et al.* 2005). With these downstream procedures performed the concentrated/diafiltered MSBD1 fraction was unable to be visualised on SDS-PAGE and did not show any antimicrobial activity against gram +/- bacteria. Several factors could be influencing this. There was not sufficient concentration peptide to allow inhibition of growth, the target of the beta-defensin may be too specific to a particular organism or virus which is unknown or that it doesn't have an antimicrobial function for example may be involved in an unknown immune function such as chemotaxis. Overall, more optimisation of this method is needed to fully ascertain a downstream process which is robust enough to fully explore the properties of reptilian beta-defensins. One such strategy which could be investigated would be when purifying the fusion MBP-MSBD1 protein from the clarified lysate by affinity chromatography instead of binding and eluting the fusion as per protocol the fusion could be bound to the column, washed with equilibration buffer, then applying the TEVp to the column and allowing to sit overnight. Then once the cleavage had occurred on the column a IMAC column could be applied in series. These columns could then be washed using the equilibration buffer as before. The TEVp would then bind to the second column allowing the liberated MSBD1 protein to be collected from the mixture.

## Critical Reflection

This doctoral journey has presented many challenges that have impacted the completion of this study. When I started this work in 2017, I joined the recently formed School of Health Sciences at Birmingham City University. During this time the university campus was going through the construction of a new purpose-built laboratory facility, however, this presented the first of a series of challenges. The new research laboratories were severely delayed, and I could not start the practical aspects of this study. Therefore, the scope of the study had to be changed to one that focused on original work but outside the scope of the laboratory environment. It was decided to follow a bioinformatic route of study with more of the emphasis on gene discovery, characterisation of the gene clusters and evolutionary analysis of the genes within these clusters. With a large amount novel information yet to be gathered, this could be attainable with open-source data and without the need for sophisticated laboratory equipment. With data freely available through the NCBI databases, newly sequenced reptilian genomes were ready for analysis. However, there was quite a drawback with undertaking these types of analyses, particularly when it came to the synonymous/nonsynonymous substitution rates and other more in-depth bioinformatic work that involved a comprehension of coding and statistical work contained in this area. This area of the research went beyond the areas of knowledge available at BCU. Fortunately, I was able to reach out to the wider scientific community, through forums I had found online and by contacting academics at other institutions, about how these programs and I was able to ascertain how they worked and the meanings of their outputs. This proved to be quite a challenge as what I had stumbled upon was quite niche and I was unfortunately unable to find confirmation of the answers I needed to give me the total confidence in my work, but not to say that the work was invalid. However, as time passed the new facility had become more complete and the design of my practical work started. By design, I incorporated work I was able to undertake due to equipment available, notably protein purification by use of the AKTA FPLC. I also wanted to design the practical work which would enhance my skills needed to enter the biotechnology sector after my studies had ended. I found at this stage that the budget for the practical undertaking was not sufficient. This required me become innovative and 'think out of the box' being as efficient as I could with the resources available to me at the time. This is something I would later come to see as an advantage and a skill much needed

and valued in the commercial sector. Working in industry reaffirms that in a development environment you are strongly encouraged to achieve as much as possible with the minimum number of resources possible.

As this work was commencing the world experienced the Covid-19 pandemic and all practical studies ceased. Back to the drawing board. Returning to the work I had initially started; the lock down gave me the chance to expand this work as it was possible to undertake this work remotely. By this point I had only analysed a few genomes and decided to investigate more species bringing the total of annotated complete gene clusters to 10, covering all 4 reptile groups. Then with the final 18 months of my studies arriving, I was able to write my thesis, explore more sequences, and when possible, try and undertake some laboratory work. Due to two extra lock downs the scope of my practical work was very limited compared to what I initially set out to do but nevertheless, I learned the skills to enable me to secure my first job working for Cobra Biologics as a development scientist, which was necessary to obtain employment due to the expiration of my registration and my bursary. Without the access to the specialist equipment at BCU I do not think I would have been successful in this application, and I am thankful for that opportunity. When my studies finished, I had the difficult time of writing up my thesis whilst trying to settle into a new and demanding job. This naturally took time and several extensions to my submission date to enable me to finally achieve this body of work, of which I am proud. I now see at the end of this journey that studying for a PhD can be a lonely furrow to plough with many ups and downs along the way. Nevertheless, I would encourage anyone with the possible opportunity to take it as this journey will enrich and develop you in more ways than you would have imagined at the outset.

## Concluding remarks and future work

Reptiles have a robust innate immune system and are evolutionarily ancient so investigating this class of animals may provide a picture of a small part of their immune properties. This work hopes to further our understanding and potentially offer new antimicrobials to help fight antimicrobial resistance. As we have seen there is a great variety of beta-defensins including charge, structure, and number. By identifying these novel, previously unknown clusters will give a foundation for further research which will provide more answers into the potential, future use of these peptides in a medical and industrial settings. Unfortunately, the work conducted in the laboratory had a few limitations which would need more work to gain a better understanding for producing and investigating these peptides. Therefore, this work provides several interesting directions future research could be conducted on. With the genomic DNA and genes within the cluster region identified, analysis of regulatory sequences may provide answers into what external responses these genes are expressed and what tissues they may be specific to. In addition, do they undergo coordinate regulation? This work may provide the basis into the evolutionary origins of these ancient genes. A more global study of the phylogeny of the reptilian beta-defensins from each of the species here could offer more understanding. Finally, optimisation of the upstream and downstream production process could allow these peptides to be produced, firstly, in quantities and speed to analyse these novel peptides but also as the basis to potentially being scaled up for commercial use.

## **9. MATERIALS AND METHODS**

### **9.1 Computer based resources**

#### **9.1.1 Multiple Sequence Alignments**

CLUSTAL X was used to produce all Multiple Sequence Alignments in this work.

#### **9.1.2 Phylogenetic trees**

Clustal Multiple Sequence Alignments in phylip format were uploaded to the IQ-tree webserver (<http://iqtree.cibiv.univie.ac.at/>) to produce trees under maximum likelihood using 1000 bootstrap replicates. The Newick output was then visualised and manipulated using FigTree V1.4.4.

#### **9.1.3 Signal peptide prediction**

Signal peptides were predicted using the SignalIP-5.0 server at <https://services.healthtech.dtu.dk/service.php?SignalP-5.0>

#### **9.1.4 Selection analysis**

To perform selection analysis on both a gene-wise and site-wise level, codon multiple sequence alignments (MSA) were needed for the downstream programs to run. To do this CLUSTAL X was used to produce an amino acid MSA. To produce this into a codon based MSA a FASTA DNA file along with the Amino acid MSA were uploaded into the form based at <http://www.bork.embl.de/pal2nal/> (Suyama *et al.* 2006).

### **9.1.5 Gene-Wise Selection analysis**

The gene-wise analysis was based on the methods described in Nei and Gojobori (1986) where the rates of synonymous ( $dS$ ) and nonsynonymous ( $dN$ ) substitutions were investigated. To produce the data the codon alignments for the signal peptide and mature/second exons were uploaded into The Synonymous and Nonsynonymous Analysis Program at <https://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html> (Korber 2000). The outputs for the proportion of ( $dS$ ) and ( $dN$ ) were plotted against each other using Excel.

### **9.1.6 Site-wise Selection Analysis**

Selection analysis of the individual amino acids within the second exon was performed using HyPhy (Pond *et al.* 2005; Pond *et al.* 2005a). The mature peptide or second exon was analysed to see which individual amino acid sites were undergoing selection. Positive and negative selection was tested by three different methods, fixed effects likelihood (FEL) (Pond *et al.* 2005b) mixed effects model of evolution (MEME) (Murrell *et al.* 2012) to detect positive selection at individual sites and fast unconstrained Bayesian approximation (FUBAR) (Murrell *et al.* 2013) MEME detects episodic positive/diversifying selection but does not detect negative/purifying selection like FEL. FUBAR can detect both episodic positive/negative selection at individual sites. The following significance levels were used for  $P \leq 0.1$  for FEL and MEME, posterior probability  $\leq 90$  for FUBAR. The sites that were undergoing positive or negative selection were plotted on an amino acid sequence logo produced using the Weblogo server (Crooks *et al.* 2004).

### **9.1.7 Genome DNA Alignment Software**

Conservation of synteny within the cluster regions were aligned using Artemis ACT: the Artemis Comparison Tool (Carver *et al.* 2005) and Mauve - multiple genome alignment (Darling *et al.* 2004).



### **9.1.8 DNA sequencing data**

All DNA sequences were obtained from the National Centre for Biotechnology Information databases.

### **9.1.9 Protein Characterisation**

Protein characterisation was undertaken using the ProtParam tool on the ExPASy server at <https://web.expasy.org/protparam/>. (Gasteiger *et al.* 2005)

## **9.2 LABORATORY PROCEDURES**

### **9.2.1 DNA Extraction Protocol for Shed Reptile Skins**

A 1-inch square piece of shed skin (cut up) was placed into 900  $\mu$ L of cell lysis buffer (10 mM Tris-base, 100 mM EDTA, 2% sodium dodecyl sulphate [SDS], pH 8.0) along with 9  $\mu$ L of proteinase K 20 mg/mL (ThermoScientific). The sample was mixed thoroughly and then placed in a dry heat block (or water bath) at 55°C, vortexed every hour for three hours and then left overnight. The sample was cooled to room temperature (Note: the skin does not dissolve after proteinase K digestion, but the DNA is released into solution) and 4  $\mu$ L of RNase A 10 mg/mL (ThermoScientific) was added, mixed and placed in a 37°C water bath for 1 h.

The sample was then cooled to room temperature again, and 300  $\mu$ L of 7.5 M ammonium acetate are added and placed on ice for 10–15 min then centrifuged at top speed (ca. 13–14k rpm) for 3 min to pellet the skin remnants, SDS and cell debris. The supernatant was drawn off and added to a microcentrifuge tube containing 900  $\mu$ L of isopropanol, this was mixed and centrifuged immediately at top speed for (16000 $\times$  g) 2 min to pellet the DNA to the bottom of the tube. After centrifugation, the isopropanol was poured off, and the pellet washed with 500  $\mu$ L of 70% ethanol. The sample was then centrifuged as before for 2 min, the ethanol was removed and then the sample placed in a vacuum concentrator (or the tube can be inverted

at room temperature for 15–20 min) until all traces of ethanol had evaporated. The DNA pellet was then resuspended in 30–100 µL of TE buffer (10 mM Tris- base, 0.1 mM EDTA, pH 8.0).

### **9.2.2 Plasmid and expression strain selection**

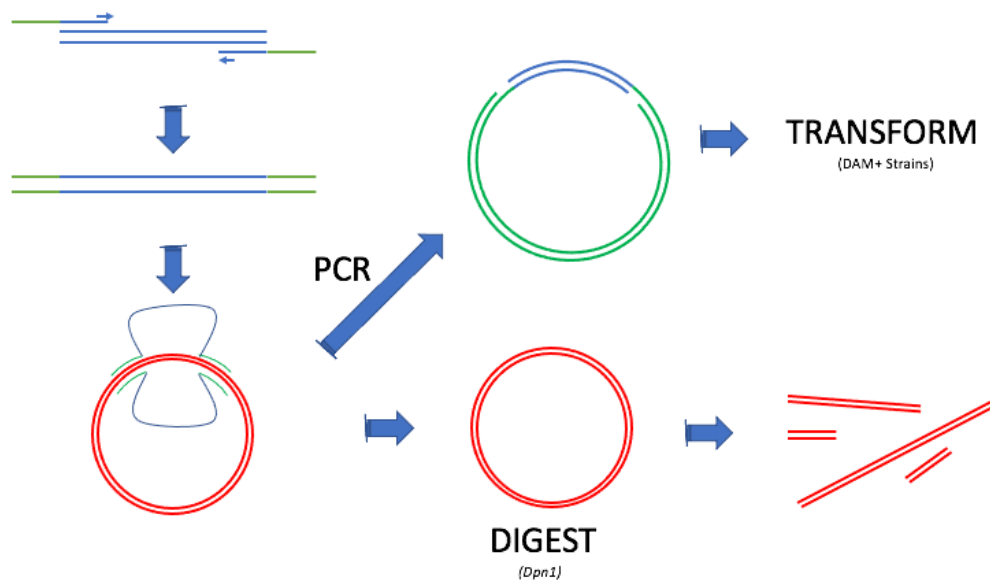
As beta-defensins are antimicrobial, a fusion protein strategy was employed. A Maltose Binding Protein (MBP) fusion vector was chosen as this has advantages of increasing solubility of the expressed protein and to alleviate potential toxic effects of expressing an antimicrobial peptide in *E. coli*. A pMAL-c6T vector was purchased from New England Biolabs. SHuffle<sup>®</sup> T7 Competent *E. coli* was purchased from New England Biolabs. This strain was chosen as it has been specifically engineered to allow the formation of disulphide bonds between cysteine residues in the cytoplasm. Beta-defensins contain 3 cystine bonds within their structure and this strain should help achieve a desired correctly folded peptide.

### **9.2.3 Restriction Free Cloning**

Restriction Free (RF) cloning is a PCR based method that allows for the insertion of a sequence into a plasmid without the need for restriction enzymes to cut the DNA and then a ligase to stitch the insert and plasmid together. It utilises large dual annealing primers in order to allow annealing to template genomic DNA for gene of interest amplification and for inserting into the plasmid multiple cloning site overcoming the limitations when the restriction sites are not present. High-fidelity PCR is first used to amplify the insert sequence from your desired template. In this case it was genomic DNA isolated from the shed python snakeskin. Once this PCR reaction was completed the fragment was then purified and used as the 'mega-primer' for use in the secondary PCR reaction using the plasmid as a template whereby the whole plasmid is amplified. When the newly synthesised PCR products anneal the mega primers act as complementary overhangs to the parental plasmid that circularise forming a nicked hybrid molecule. Dpn1 restriction enzyme (Thermoscientific) was used to degrade the methylated parental plasmid and leaving behind the unmethylated daughter plasmid with the insert intact and ready to transform directly into the competent bacteria. Transformation has to take place in a DAM+ bacterial strain as the Dpn1 is used to selectively digest the parental

DNA after the second reaction therefore purification will produce methylated plasmids once grown out (figure 9.1)

Hybrid primers were designed (figure 9.2 and section 9.2.4) to facilitate RF cloning of the python exon. Forward and reverse primers were located at each end of the desired region of the python second exon making sure the codons were aligned with the sequence to keep the reading frame and to also introduce a stop codon at the end. As well as this the primer pair for insertion of the Multiple Cloning Site (MCS) was established. These primers were then put into a T<sub>m</sub> calculator to ensure annealing temperatures were correct, according to best practices (Dieffenbach *et al* 1993, Angelica and Fong 2010). It is recommended that the side of the primers for the insert be at 55°C and 60°C for the side of primer for the plasmid.



**Figure 9.1 Schematic of the RF cloning protocol.**

Hybrid primers are designed with complementary sequences to insert (Blue) and plasmid (green) and a first round of PCR is then performed to produce a 'mega-primer' with the insert sequence flanked by sequences complementary to the desired position in the MCS. A second round of PCR is performed whereby the mega-primer initiates replication of the parental plasmid (Red). Since the parental plasmid is replicated, mega-primer binding on the daughter plasmid fails to expose the 3'-end for elongation and therefore a linear product is produced. This daughter plasmid is then transformed into DAM+ strains for purification before being in downstream processes.

A. Plasmid Sequence – Plasmid pMAL-c6T Vector (New England Biolabs)

```

2641 AGACTAATTCGAGCTCGAACCAACAACAATAACAATAACAACAACCTCGGGGAGAACC 2700
----:----|----:----|----:----|----:----|----:----|----:----|
2641 TCTGATTAAGCTCGAGCTTGTGTGTTGTTGTTATTGTTATTGTTGTTGGAGCCCCCTCTTGG 2700

2701 TGTACTTCCAGATGCTGATGGGCGGCCGCGATATCGTTCGACGGATCCGAATTCCTGCAG 2760
----:----|----:----|----:----|----:----|----:----|----:----|
2701 ACATGAAGGTCTACGACTACCCGCCGGCGCTATAGCAGCTGCCTAGGCTTAAGGGACGTC 2760

2761 CTAATTAAATAAGCTTCAAATAAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTT 2820
----:----|----:----|----:----|----:----|----:----|----:----|
2761 CATTAAATTTATTCGAAGTTTATTTTGCTTCCGAGTCAGCTTCTGACCCGAAAGCAA 2820

```

B. Insert sequence from Python *Molurus Bivittatus* – Accession no NW\_006538925.1

```

S T A A Q S S F H L Q Q R C R P L V Q V F1
P Q P H R V P F I C S S V A G H W C K C F2
H S R T E F L S F A A A L P A I G A S A F3
10561 TCCACAGCCGCACAGAGTTCCTTTCATTTCAGCAGCGTTGCCGCCATTGGTGCAAGTG 10620
----:----|----:----|----:----|----:----|----:----|----:----|
10561 AGGTGTCGGCGTGTCTCAAGGAAAGTAAACGTCGTCGCAACGGCCGGTAACCACGTTTAC 10620
E V A A C L E K * K C C R Q R G N T C T F6
N W L R V S N R E N A A A N G A M P A L F5
G C G C L T G K M Q L L T A P W Q H L H F4

P I I C S W W K H S L R Q C P * L L W H F1
R L S V H G G N T V S G S V H D C C G T F2
D Y L F M V E T Q S P A V S M I V V A H F3
10621 CCGATTATCTGTTTCATGGTGGAAACACAGTCTCCGGCAGTGTCCATGATTGTTGTGGCAC 10680
----:----|----:----|----:----|----:----|----:----|----:----|
10621 GGCTAATAGACAAGTACCACCTTTGTGTCAGAGGCCGTACAGGTACTAACACACCGTG 10680
G I I Q E H H F C L R R C H G H N N H C F6
A S * R N M T S V C D G A T D M I T T A F5
R N D T * P P F V T E P L T W S Q Q P V F4

T R L S * R S P F A I Q R K R F L I S L F1
L G C H K G P P L L F R E S G F * L V W F2
* A V I K V P L C Y S E K A V F N * F G F3
10681 ACTAGGCTGTCATAAAGGTCCCCCTTGTCTATTCAGAGAAAGCGGTTTTTAATTAGTTTG 10740
----:----|----:----|----:----|----:----|----:----|----:----|
10681 TGATCCGACAGTATTTCCAGGGGAAACGATAAGTCTCTTTCGCCAAAAATTAATCAAAC 10740
V L S D Y L D G K A I * L F R N K I L K F6
C * A T M F T G R Q * E S F A T K L * N F5
S P Q * L P G G K S N L S L P K * N T Q F4

```

**Figure 9.2 Translation map of sequences for RF Cloning.**

A) Plasmid sequence showing multiple cloning region (red) and part of primers used from mega primer (blue). B) Second exon of Python Beta-defensin (yellow) and part of primers (green).

### 9.2.4 Primers

Forward Primer

Plasmid annealing = 60°C Target annealing = 55°C Length = 42

CGGGGAGAACCTGTACTIONCCAG AAGGGGGACCTTTATGACAG

Reverse Primer

Plasmid annealing = 59°C Target annealing = 55°C Length = 49

TGAAGCTTATTTAATTACCTGCAGGGGAATTC TCATTTGCAGCAGCGTTG

*Hybrid primers showing both regions and the sequences used.*

### 9.2.5 PCR Reaction Protocols

The Insert reaction was set up by mixing 20µl Nuclease Free Water, 10µl 2x PCR Master mix (Phusion Flash High-Fidelity, Thermo Scientific), 0.5µM Hybrid Primers (MERCK) and 2µl Template DNA (Python Snakeskin). The reaction conditions were an initial denaturation cycle at 98°C for 2 minutes, an annealing cycle at 50°C for 30 seconds, an extension cycle at 72°C for 10 seconds and denaturing cycle at 98°C for 10 seconds a total of 40x cycles. Plasmid reaction was set up by the mixing of 50µl Nuclease Free Water, 25µl of 2x PCR Master mix (Phusion Flash High-Fidelity, Thermo Scientific), 5µl – PCR product from insert reaction 2.5µl of pMAL-c6T Vector (10µg (50µl) – (New England Biolabs). Reaction Conditions were an Initial denaturation cycle at 98°C for 30 seconds, an annealing cycle at 56°C for 30 seconds an extension cycle at 72°C for 1 minute and 45 seconds and a denaturing cycle at 98°C for 30 seconds for a total of 35x cycles.

PCR plasmid products were then digested with 1µl of Dpn1 (10U/µl) (ThermoFisher Scientific) for 2 hours at 37°C. This was then transformed into DH5α *E. coli* (New England Biolabs).

### **9.2.6 Agarose Gel Electrophoresis**

Agarose gel electrophoresis was performed using Agarose (Sigma) at either 3% or 0.8% agarose gels for checking the inserts on the vector and plasmid size respectively. To produce either agarose solution % w/v was utilised. For example, 3% gel was made up by adding 3g of agarose to 100ml of 1X TBE buffer made from 10X TBE buffer (National Diagnostics). 5ul of Midori Green Advance DNA stain (Nippon Genetics Europe GmbH) was added to 100ml of preheated and dissolved agarose solution and mixed gently. Once the temperature was around 70-60°C, this was cast into a gel tray with a 20 well comb to a thickness of around 0.5cm. A MultiSUB Midi Horizontal Gel Unit with 10 x 7cm tray was run at 100V for between 30-40 minutes. The ladders that were used for sizing were GeneRuler Low Range DNA ladder (Thermoscientific) and 1KbPLUS DNA ladder (Geneflow) for plasmid insert and plasmid sizing respectively. 6x Loading dye (Thermoscientific) was added to the samples to track migration. Visualisation was achieved in a G: BOX Chemi XX9 Gel Imaging System using UV at 254nm and system software for image manipulation.

### **9.2.7 Protocol for preparation of electrocompetent *E. coli* for transformation of plasmid in to holding strain DH5 $\alpha$**

This method was adapted from Gonzales *et al* (2013)

Preparation of Bacterial Cultures, Tools, and Reagents was performed during the afternoon of day 1. 5 ml autoclaved LB broth in sterile borosilicate glass test tubes was inoculated with a small aliquot of *E. coli* DH5 $\alpha$  (New England Biolabs) and placed in a shaking incubator housed at 37 °C at 180rpm overnight. LB-agar (Sigma Aldrich) plates with and without Ampicillin prepared and ddH<sub>2</sub>O was autoclaved and all stored at 4 °C. The following morning 100  $\mu$ l of the O/N bacterial culture was spread onto each LB agar plate and incubated at 37 °C for 4-6 hr, or until a thin lawn of bacterial growth was distinguishable. The bacteria were harvested with a sterile inoculating loop making sure not to pierce or break the surface of the agar. One 2 mm diameter bacterial mass was for transformation. This was then resuspended in 1 ml ice-cold sterile ddH<sub>2</sub>O, mixed well until no clumps were visible and kept on ice. The bacterial suspension was centrifuged for 5 min at 5,000 x g in a refrigerated microcentrifuge set to 4

°C. The supernatant was discarded, and the bacterial pellet was resuspended in the same volume of sterile ddH<sub>2</sub>O. This was repeated twice more for a total of three washes. After the final wash the supernatant was resuspended in 40 µl ice cold sterile ddH<sub>2</sub>O and kept on ice. 2µl of cloned plasmid DNA (pMAL-c6T) was added to the 40µl bacterial suspension and transferred into a pre- chilled, sterile 0.1 cm gap cuvette. The salt concentration in the DNA sample must be low, as it will contribute to arcing of the pulse in the next step. Electroporation was performed (Mirus Bio - Ingenio<sup>®</sup> EZporator<sup>®</sup> Electroporation System) at 1250V The time constant should be ~5.0 msec, and no arcing should occur. The cell suspension was quickly recovered by resuspending into 1 ml LB broth and transferring into previously autoclaved bijoux. The cells were allowed to recover by incubating under aerated growth conditions in shaker at 37 °C for 30 min without ampicillin; 100µl of bacteria were placed onto the previously prepared LB agar plates with ampicillin and incubated at 37 °C overnight.

### ***9.2.8 Recombinant plasmid purification from holding strain for transformation into SHuffle<sup>®</sup> Competent E. coli***

Thermo Scientific GeneJET plasmid miniprep kit was utilised for this step to purify the recombinant plasmid ready for transformation into expression strain, following manufacturer's protocol. All steps were carried out at room temperature and all centrifugations were carried out in a microcentrifuge at  $\geq 12\ 000 \times g$  (10 000-14 000 rpm, depending on the rotor type). The pelleted cells from an overnight culture were resuspended in 250 µL of Resuspension Solution and vortexed, followed by the addition 250 µL of Lysis Solution and invert the tube 4-6 times. 350 µL of Neutralization Solution was added and the tube inverted 4-6 times. This mixture was centrifuged for 5 minutes. The supernatant was transferred to a Thermo Scientific GeneJET Spin Column and centrifuged for 1 minute. 500 µL of Wash Solution was added and centrifuged for 30-60 s. this was repeated 2 times with the flowthrough discarded each time with a final centrifuge of the empty column for 1 minute. The column was transferred to a new tube and 50 µL of Elution Buffer was added to the column and incubated for 2 minutes then Centrifuged for 2 minutes to collect the flow-through containing the pDNA.

### **9.2.9 Plasmid Preparation into SHuffle® Competent *E. coli*.**

SHuffle® Competent *E. coli* (New England Biolabs) was used for protein expression as this strain has been specifically Engineered *E. coli* K12 to promote disulphide bond formation in the cytoplasm. The *E. coli* was transformed following the manufacturer's protocol. A tube of competent *E. coli* cells was thawed on ice for 10 minutes. 1-5µl containing 1 pg-100 ng of plasmid DNA was added to the cell mixture (without vortexing). The mixture was placed on ice for 30 minutes and then heat shocked at 42°C for 30 seconds and then placed on ice for 5 minutes. All stages were without mixing. 250µl of room temperature SOC media was pipetted into the mixture and shaken at 250rpm at 37°C for 60 minutes. Whilst the shaking was being performed ampicillin selection plates were warmed to 37°C. The cells were then thoroughly mixed by flicking the tube and inverting; then several 10-fold serial dilutions in SOC media were carried out. 50-100µl of each dilution were spread onto a selection plate and incubated overnight at 37°C.

### **9.2.10 DNA sequencing of transformed *E. coli***

DNA sequencing was outsourced to The University of Birmingham DNA services and performed on a capillary sequencer ABI 3730 (Applied Biosystems).

### **9.2.11 Expression of Python Beta-defensin in SHuffle® *E. coli*.**

Following transformation of expression strain the transformed cells were grown overnight in 10ml LB medium (10 g/L Bacto-tryptone, 5 g/L yeast extract, 10 g/L NaCl) containing 100µg/ml ampicillin under shaking conditions at 37°C. 1% (v/v) of overnight culture was inoculated into 1L of fresh LB medium and was grown until the OD<sub>600</sub> reached 0.3 at which point the culture was induced with IPTG (Sigma Aldrich) at a concentration of 1mM IPTG. 4h post induction the cells were harvested by centrifugation at 4000 x g for 20 minutes at 4°C.



### ***9.2.12 Cell lysis for Beta-defensin recovery***

The harvested cells were resuspended in 25ml of lysis buffer consisting of the equilibration buffer for purification (50mM Tris-HCl, 200mM NaCl, pH7.4) (Sigma Aldrich) supplemented with SIGMAFAST™ Protease Inhibitor Cocktail Tablets, EDTA-Free. Once resuspended the cells were lysed with sonication using a 6mm probe attached to Sonics Vibra Cell VCX505 at 40% power. The Cells were placed on ice to reduce heat and cycled for 5 seconds on 10 seconds off for a total of 10 minutes. Once lysed the supernatant was recovered by centrifugation at 4000 x g for 30 minutes. The supernatant was recovered and filtered through a 0.2µm syringe filter ready for purification.

### ***9.2.13 Solubility Testing for expressed MSBD1***

A small aliquot was taken from the post sonication mixture and was centrifuged to pellet the cell debris. The supernatant was drawn off and the cell pellet was resuspended in ddH<sub>2</sub>O. 20µl of each sample was added to SDS-PAGE sample buffer and heated for 10mins at 95°C. These aliquots were then loaded onto a 12% SDS-PAGE gel and run at 0.3mA for 80 minutes

#### **9.2.14 Affinity Purification of fusion MPB-MSBD1 fusion protein**

An ÄKTA Pure 25 purification system was employed for the purification of the fusion protein. Affinity chromatography column used for this stage was a 1ml MBPTrap HP (Cytiva). 5ml of supernatant was loaded onto the column using a 5ml loop.

#### *Chromatography operating conditions*

	<b>FLOWRATE (ml/min)</b>	<b>COLUMN VOLUMES (CV)</b>	<b>BUFFER</b>
EQUILIBRATION	1	10	50mM Tris-HCl, 200mM NaCl, pH7.4
SAMPLE LOAD	1	5	50mM Tris-HCl, 200mM NaCl, pH7.4
WASH	1	5	50mM Tris-HCl, 200mM NaCl, pH7.4
ELUTION	1	5 of 100% elution buffer	50mM Tris-HCl, 200mM NaCl, 10mM Maltose, pH7.4

\*At all stages the fractions were collected for post analysis using SDS-PAGE.

#### **9.2.15 SDS-PAGE analysis**

All SDS-PAGE analysis was performed using the SureCast™ Handcast System (Invitrogen). Following the manufacturers protocol a mixture of 10%, 20% and 12-20% gradient gels were cast and run at 0.3mA for 80 minutes. All gels were the stained using a standard Coomassie Blue Stain Protocol. The running buffer used was 3g Tris base (Sigma Aldrich), 14.4g Glycine (Sigma Aldrich), 10g SDS (Sigma Aldrich) per litre and the loading buffer (4x) was 2ml 1M Tris-HCl pH 6.8, 0.8 g SDS.4.0 ml 100% glycerol, 0.4 ml 14.7 M β-mercaptoethanol, 1.0 ml 0.5 M EDTA, 8 mg bromophenol Blue (Sigma Aldrich). The protein Ladders used were PageRuler Plus prestained protein ladder, 10-250kDa (ThermoFisher) and PageRuler Unstained Low

Range Ladder (ThermoFisher). The Coomassie staining reagents were made as follows; Coomassie Blue stain was produced by dissolving 0.4g of Coomassie blue R350 in 200 mL of 40% (v/v) HPLC grade methanol in water with stirring as needed. The solution was filtered to remove any insoluble material, and 200mL of 20% (v/v) acetic acid in water added. The final concentration is 0.1% (w/v) Coomassie blue R350, 20% (v/v) methanol, and 10% (v/v) acetic acid. Gel Fixing Solution was made by adding 500mL of USP-grade 95% (v/v) ethanol to 300 mL of HPLC grade water. 100 mL of reagent grade acetic acid was added and the total volume adjusted to 1000 mL with water. The final concentrations are 50% (v/v) ethanol in water with 10% (v/v) acetic acid. Gel Destaining Solution was made by the addition of 500mL of HPLC-grade methanol to 300 mL of HPLC grade water. 100 mL of reagent grade acetic acid was added and, after mixing, the total volume was adjusted to 1000mL with water. The final concentrations are 50% (v/v) methanol in water with 10% (v/v) acetic acid and Storage Solution was 25mL of reagent grade acetic acid to 400mL of HPLC grade water. After mixing, the final volume was adjusted to 500mL with water. The final concentration of acetic acid is 5% (v/v).

#### ***9.2.16 Staining Procedure***

Once the gel had been run for the required time it was removed from the glass plate and placed into the fixing solution for 1h with gentle agitation to remove the running buffer from the gel and to fix the proteins. Then the gel was transferred into the Coomassie stain solution and left to shake gently overnight. After staining the gel was removed and placed into the gel destain solution and gently agitated until the background of the gel had been completely removed and the protein bands were visible. Gels were then stored in the storage solution until they were visualised using a G: BOX Chemi XX9 Gel Imaging System under visible light.

### **9.2.17 Cleavage of MBP fusion protein to release MSBD1**

Tobacco Etch Virus (TEV) protease tagged with poly-histidine was purchased from New England Biolabs and used to cleave the fusion protein. 10 µl of TEV protease was added to 1 ml of purified MBP-PBBD1 fusion protein and incubated overnight at 30°C to release the PBBD1 peptide.

### **9.2.18 Immobilised metal ion affinity (IMAC) chromatography for purification of MSBD1.**

Both the MBP tag and TEV protease had N terminal poly-histidine tags to allow simple purification of PBBD1 using immobilised metal ion affinity (IMAC) chromatography. This allowed the cleaved PBBD1 to be collected in the flowthrough. The column that was used for the IMAC chromatography was HisTrap™ HP 1 ml column (Cytiva). 5 ml of post cleavage mixture was loaded onto the column.

#### *Operating Conditions*

	<b>FLOWRATE (ml/min)</b>	<b>COLUMN VOLUMES (CV)</b>	<b>BUFFER</b>
EQUILIBRATION	1	10	20 mM sodium phosphate, 0.5 M NaCl, 5 mM imidazole, pH 7.4
SAMPLE LOAD	1	5	50mM Tris-HCl, 200mM NaCl, pH7.4
WASH	1	5	20 mM sodium phosphate, 0.5 M NaCl, 5 mM imidazole, pH 7.4
ELUTION	1	5CV of 100% elution buffer	20 mM sodium phosphate, 0.5 M NaCl, 0.5 M imidazole, pH 7.4

\*At all stages the fractions were collected for post analysis using SDS-PAGE.

### 9.2.19 Refolding and purification of PBBD1 by Size Exclusion Chromatography

The denaturation of the post cleavage mixture was achieved by putting the post cleavage mixture in 8M Urea (Sigma Aldrich) and 10mM DDT (Sigma Aldrich) for 2H at room temperature.

The column used in this purification strategy was HiLoad® 16/600 Superdex® 75 pg (Cytiva). 5ml of the denatured post cleavage mixture was added as load material and 5ml fractions were collected throughout the run.

#### Operating Conditions

	<b>FLOWRATE (ml/min)</b>	<b>COLUMN VOLUMES (CV)</b>	<b>REFOLDING BUFFER</b>
EQUILIBRATION	1	1	2M urea, 0.5M arginine, 150mM NaCl, pH 8.0, 1mM GSH, 1mM GSSG
SAMPLE LOAD	1	1.5	2M urea, 0.5M arginine, 150mM NaCl, pH 8.0, 1mM GSH, 1mM GSSG

### 9.2.20 Preliminary Antimicrobial Activity testing

All Antimicrobial testing was performed using Muller-Hinton agar plates. Bacterial strains used in the study were *E. coli* and *B. cereus*.

### 9.2.21 Diafiltration

Diafiltration was performed using Sartorius Vivaspin 20, 3000kDa MWCO PES ultrafiltration units. Exchange buffer was 50mM Tris-HCL, 25mM NaCl. Buffer exchange was performed by centrifugation at 3500xg and diafiltered 3 times in exchange buffer.

## References

1. A.F.A. Smit, R. Hubley & P. Green, unpublished data. Current Version: open-4.0.9, <http://www.repeatmasker.org/>
2. Akashi, H. (1995) 'Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA', *Genetics*, 139(2), pp. 1067–1076. doi: 10.1093/genetics/139.2.1067.
3. Akira, S., Uematsu, S. and Takeuchi, O. (2006) 'Pathogen recognition and innate immunity', *Cell*, 124(4), pp. 783–801. doi: 10.1016/j.cell.2006.02.015.
4. Alibardi L, Celegghin A, Dalla Valle L, (2012) Wounding in lizards results in the release of beta-defensins at the wound site and formation of an anti-microbial barrier. *Dev. Comp. Immunol*, **36**, 557-565
5. Alibardi L, (2010) Ultrastructural features of the process of wound healing after tail and limb amputation in Lizard. *Acta Zool*, **91**, 306-318.
6. Allen, I. C., Moore, C. B., Schneider, M., Lei, Y., Davis, B. K., Scull, A., Gris, D., Roney, K. E., Zimmermann, A. G., Bowzard, J. B., Monroe, K. M., Pickles, R. J., Sambhara, S., Ting, P. Y., & Hill, C. (2012). NLRX1 protein attenuates inflammatory responses to virus infection by interfering with the RIG-I-MAVS signalling pathway and TRAF6 ubiquitin ligase. *Immunity*, 34(6), 854–865. <https://doi.org/10.1016/j.immuni.2011.03.026>.NLRX1
7. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*. 2019;37(4):420-423. doi:10.1038/s41587-019-0036-z
8. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Altschul et al. 1990. Basic Local Alignment Search Tool.pdf. In *Journal of Molecular Biology* (Vol. 215, Issue 3, pp. 403–410). [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
9. Anders, H. J. (2007). Innate pathogen recognition in the kidney: Toll-like receptors, NOD-like receptors, and RIG-like helicases. In *Kidney International* (Vol. 72, Issue 9, pp. 1051–1056). <https://doi.org/10.1038/sj.ki.5002436>
10. Angelica, M. D. and Fong, Y. (2010) 'Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids', *Biotechniques*, 48(6), pp. 463–465. doi: 10.2144/000113418.Overlap.
11. Aoki, W. and Ueda, M. (2013) 'Characterization of antimicrobial peptides toward the development of novel antibiotics', *Pharmaceuticals*, 6(8), pp. 1055–1081. doi: 10.3390/ph6081055.

12. Appendini, P. and Hotchkiss, J. H. (2002) 'Review of antimicrobial food packaging', *Innovative Food Science and Emerging Technologies*, 3(2), pp. 113–126. doi: 10.1016/S1466-8564(02)00012-7.
13. Ausubel, F. M. (2005) 'Are innate immune signaling pathways in plants and animals conserved?', *Nature Immunology*, 6(10), pp. 973–979. doi: 10.1038/ni1253.
14. Bagheri, M., Beyermann, M. and Dathe, M. (2009) 'Immobilization reduces the activity of surface-bound cationic antimicrobial peptides with no influence upon the activity spectrum', *Antimicrobial Agents and Chemotherapy*, 53(3), pp. 1132–1141. doi: 10.1128/AAC.01254-08.
15. Bai, L. L., Yin, W. B., Chen, Y. H., Niu, L. L., Sun, Y. R., Zhao, S. M., Yang, F. Q., Wang, R. R. C., Wu, Q., Zhang, X. Q. and Hu, Z. M. (2013) 'A New Strategy to Produce a Defensin: Stable Production of Mutated NP-1 in Nitrate Reductase-Deficient *Chlorella ellipsoidea*', *PLoS ONE*, 8(1). doi: 10.1371/journal.pone.0054966.
16. Bals R. (2000). Antimikrobielle Peptide und Peptidantibiotika [Antimicrobial peptides and peptide antibiotics]. *Medizinische Klinik (Munich, Germany: 1983)*, 95(9), 496–502. <https://doi.org/10.1007/pl00002139>
17. Benato, F., Dalla Valle, L., Skobo, T., & Alibardi, L. (2013). Biomolecular Identification of Beta-defensin-Like Peptides from the Skin of the Soft-Shelled Turtle *Apalone spinifera*. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 320(4), 210–217. <https://doi.org/10.1002/jez.b.22495>
18. Biémont, C., & Vieira, C. (2006). Genetics: junk DNA as an evolutionary force. *Nature*, 443(7111), 521–524. <https://doi.org/10.1038/443521a>
19. Biragyn, A., Ruffini, P. A., Leifer, C. A., Klyushnenkova, E., Shakhov, A., Chertov, O., Shirakawa, A. K., Farber, J. M., Segal, D. M., Oppenheim, J. J., & Kwak, L. W. (2002). Toll-like receptor 4-dependent activation of dendritic cells by  $\beta$ -defensin 2. *Science*, 298(5595), 1025–1029. <https://doi.org/10.1126/science.1075565>
20. Bulet, P., Stocklin, R., and Menin, L. (2004) Anti-microbial peptides: from invertebrates to vertebrates. *Immunological Reviews*, **198**, 169–184.
21. Burge, C. and Karlin, S. (1997) 'Prediction of complete gene structures in human genomic DNA' Edited by F. E. Cohen', *Journal of Molecular Biology*, 268(1), pp. 78–94. Available at: <http://www.sciencedirect.com/science/article/pii/S0022283697909517>.
22. Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., & Parkhill, J. (2005). ACT: The Artemis comparison tool. *Bioinformatics*, 21(16), 3422–3423. <https://doi.org/10.1093/bioinformatics/bti553>
23. Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based

experimental data. *Bioinformatics*, 28(4), 464–469.  
<https://doi.org/10.1093/bioinformatics/btr703>

24. Chattopadhyay, S., Sinha, N. K., Banerjee, S., Roy, D., Chattopadhyay, D., & Roy, S. (2006). Small cationic protein from a marine turtle has beta-defensin-like fold and antibacterial and antiviral activity. *Proteins*, 64(2), 524–531.  
<https://doi.org/10.1002/prot.20963>

25. Chen, G., Shaw, M. H., Kim, Y. G., & Nuñez, G. (2009). NOD-like receptors: Role in innate immunity and inflammatory disease. *Annual Review of Pathology: Mechanisms of Disease*, 4, 365–398. <https://doi.org/10.1146/annurev.pathol.4.110807.092239>

26. Chan, Y. S. and Ng, T. B. (2013) 'Northeast Red Beans Produce a Thermostable and pH-Stable Defensin-Like Peptide with Potent Antifungal Activity', *Cell Biochemistry and Biophysics*. Springer US, 66(3), pp. 637–648. doi: 10.1007/s12013-012-9508-1.

27. Chen, H., Ma, M. Y., Sun, L., Fang, S. G., & Wan, Q. H. (2015). Genomic structure and evolution of beta-defensin genes in the golden pheasant and hwamei. *Science Bulletin*, 60(7), 679–690. <https://doi.org/10.1007/s11434-015-0758-3>

28. Chen, J., Shang, S., Wu, X., Zhong, H., Zhao, C., Wei, Q., Zhang, H., Xia, T., Chen, Y., Zhang, H., & Tang, X. (2019). Genomic analysis and adaptive evolution of the RIG-I-like and NOD-like receptors in reptiles. *International Journal of Biological Macromolecules*, 134, 1045–1051. <https://doi.org/10.1016/j.ijbiomac.2019.05.172>

29. Cheng, Y., Prickett, M.D., Gutowska, W. *et al.* Evolution of the avian  $\beta$ -defensin and cathelicidin genes. *BMC Evol Biol* 15, 188 (2015). <https://doi.org/10.1186/s12862-015-0465-3>

30. Chromek M, Arvidsson I, Karpman D (2012) The antimicrobial peptide cathelicidin protects mice from Escherichia coli O157:H7-mediated disease, *Plos One*, 7, e46476

31. Conant, G. C. and Wolfe, K. H. (2008) 'Turning a hobby into a job: How duplicated genes find new functions', *Nature Reviews Genetics*, 9(12), pp. 938–950. doi: 10.1038/nrg2482.

32. Contreras, G., Shirdel, I., Braun, M. S., & Wink, M. (2020). Defensins: Transcriptional regulation and function beyond antimicrobial activity. *Developmental and Comparative Immunology*, 104(August 2019), 103556. <https://doi.org/10.1016/j.dci.2019.103556>

33. Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature reviews. Genetics*, 10(10), 691–703. <https://doi.org/10.1038/nrg2640>

34. Correa, P. G. and Oguiura, N. (2013) 'Phylogenetic analysis of  $\beta$ -defensin-like genes of Bothrops, Crotalus and Lachesis snakes', *Toxicon*, 69, pp. 65–74. doi: 10.1016/j.toxicon.2013.02.013.



35. Creagh, E. M. and O'Neill, L. A. J. (2006) 'TLRs, NLRs and RLRs: a trinity of pathogen sensors that co-operate in innate immunity', *Trends in Immunology*, 27(8), pp. 352–357. doi: 10.1016/j.it.2006.06.003.
36. Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome research*, 14(6), 1188–1190. <https://doi.org/10.1101/gr.849004>
37. Crovella, S., Antcheva, N., Zelezetsky, I., Boniotto, M., Pacor, S., Verga Falzacappa, M. V., & Tossi, A. (2005). Primate beta-defensins--structure, function and evolution. *Current Protein & Peptide Science*, 6, 7–21.
38. Dalla Valle, L., Benato, F., Maistro, S., Quinzani, S., & Alibardi, L. (2012). Bioinformatic and molecular characterization of beta-defensins-like peptides isolated from the green lizard *Anolis carolinensis*. *Developmental and Comparative Immunology*, 36(1), 222–229. <https://doi.org/10.1016/j.dci.2011.05.004>
39. Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7), 1394–1403. <https://doi.org/10.1101/gr.2289704>
40. Derache, C., Labas, V., Aucagne, V., Meudal, H., Landon, C., Delmas, A. F., Magallon, T., & Lalmanach, A. C. (2009). Primary structure and antibacterial activity of chicken bone marrow-derived  $\beta$ -defensins. *Antimicrobial Agents and Chemotherapy*, 53(11), 4647–4655. <https://doi.org/10.1128/AAC.00301-09>
41. Deslouches B, Phadke SM, Lazarevic V, Cascio M, Islam K, Montelaro RC, et al. De novo generation of cationic antimicrobial peptides: influence of length and tryptophan substitution on antimicrobial activity. *Antimicrob Agents Chemother*. 2005; 49:316–22.
42. Diamond, G., Kaiser, V., Rhodes, J., Russell, J. P., & Bevins, C. L. (2000). Transcriptional regulation of  $\beta$ -defensin gene expression in tracheal epithelial cells. *Infection and Immunity*, 68(1), 113–119. <https://doi.org/10.1128/IAI.68.1.113-119.2000>
43. Dieffenbach, C. W., Lowe, T. M. J. and Dveksler, G. S. (1993) 'General concepts for PCR primer design', *Genome Research*, 3(3). doi: 10.1101/gr.3.3. S30.
44. Eddy, S. (1998) 'Profile hidden Markov models.', *Bioinformatics*, 14(9), pp. 755–763. doi: btb114 [pii].
45. Edwards, T., Karl, A. E., Vaughn, M., Rosen, P. C., Torres, C. M., & Murphy, R. W. (2016). The desert tortoise trichotomy: Mexico hosts a third, new sister-species of tortoise in the gopherus morafkai–G. agassizii group. *ZooKeys*, 2016(562), 131–158. <https://doi.org/10.3897/zookeys.562.6124>

46. Eirín-López, J. M., Rebordinos, L., Rooney, A. P., & Rozas, J. (2012). The birth-and-death evolution of multigene families revisited. *Genome Dynamics*, 7, 170–196. <https://doi.org/10.1159/000337119>
47. Emes RD, Goodstadt L, Winter EE, Ponting CP (2003), Comparison of the genomes of human and mouse lays the foundation of genome zoology *Hum Mol Genet.* 2003 Apr 1;12(7):701-9.
48. Etienne O, Picart C, Taddei C, et al. Multilayer Polyelectrolyte Films Functionalized by Insertion of Defensin: A New Approach to Protection of Implants from Bacterial Colonization. *Antimicrobial Agents and Chemotherapy.* 2004;48(10):3662-3669. doi:10.1128/AAC.48.10.3662-3669.2004.
49. Erwin, D.H. and Davidson, E.H. (2002) The last common bilaterian ancestor. *Development*, **129**, 3021–3032.
50. Evans, E. W., Beach, F. G., Moore, K. M., Jackwood, M. W., Glisson, J. R., & Harmon, B. G. (1995). Antimicrobial activity of chicken and turkey heterophil peptides CHP1, CHP2, THP1, and THP3. *Veterinary Microbiology*, 47(3–4), 295–303. [https://doi.org/10.1016/0378-1135\(95\)00126-3](https://doi.org/10.1016/0378-1135(95)00126-3)
51. Felgueiras, H. P. and Amorim, M. T. P. (2017) ‘Functionalization of electrospun polymeric wound dressings with antimicrobial peptides’, *Colloids and Surfaces B: Biointerfaces*. Elsevier B.V., 156, pp. 133–148. doi: 10.1016/j.colsurfb.2017.05.001.
52. Ferraro, J. V. and Binetti, K. M. (2014) ‘American alligator proximal pedal phalanges resemble human finger bones: Diagnostic criteria for forensic investigators’, *Forensic Science International*. Elsevier Ireland Ltd, 240, pp. 151.e1-151.e7. doi: 10.1016/j.forsciint.2014.04.011.
53. Fetzner, J. W. (1999) ‘Extracting high-quality DNA from shed reptile skins: A simplified method’, *BioTechniques*, 26(6), pp. 1052–1054.
54. Fiston-Lavier, A. S., Anxolabehere, D. and Quesneville, H. (2007) ‘A model of segmental duplication formation in *Drosophila melanogaster*’, *Genome Research*, 17(10), pp. 1458–1470. doi: 10.1101/gr.6208307.
55. Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., & Mathelier, A. (2020). JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Research*, 48(D1), D87–D92. <https://doi.org/10.1093/nar/gkz1001>
56. Froy, O. and Gurevitz, M. (2003) ‘Arthropod and mollusk defensins - Evolution by exon-shuffling’, *Trends in Genetics*, 19(12), pp. 684–687. doi: 10.1016/j.tig.2003.10.010.

57. Funderburg N, Lederman MM, Feng Z, Drage MG, Jadowsky J, Harding CV, Weinberg A, Sieg SF (2007) Human  $\alpha$ -defensin-3 activates professional antigen-presenting cells via Toll-like receptors 1 and 2. *Proc Natl Acad Sci USA* 104(47):18631– 18635
58. Ganz T, Selsted ME, Szklarek D, Harwig SS, Daher K, Bainton DF, Lehrer RI (1985) Defensins. Natural peptide antibiotics of human neutrophils. *J Clin Invest* 76, 1472-1435
59. Ganz, T. (2003) 'Defensins: Antimicrobial peptides of innate immunity', *Nature Reviews Immunology*, 3(9), pp. 710–720. doi: 10.1038/nri1180
60. Ganz, T. (2004) 'Defensins: antimicrobial peptides of vertebrates', *Comptes Rendus Biologies*, 327(6), pp. 539–549. doi: <https://doi.org/10.1016/j.crvi.2003.12.007>.
61. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). The Proteomics Protocols Handbook. *The Proteomics Protocols Handbook*, 571–608. <https://doi.org/10.1385/1592598900>
62. Girardin, S. E., Tournebise, R., Mavris, M., Page, A. L., Li, X., Stark, G. R., Bertin, J., Distefano, P. S., Yaniv, M., Sansonetti, P. J., & Philpott, D. J. (2001). CARD4/Nod1 mediates NF- $\kappa$ B and JNK activation by invasive *Shigella flexneri*. *EMBO Reports*, 2(8), 736–742. <https://doi.org/10.1093/embo-reports/kve155>.
63. Goddard, J. M. and Hotchkiss, J. H. (2007) 'Polymer surface modification for the attachment of bioactive compounds', *Progress in Polymer Science (Oxford)*, 32(7), pp. 698–725. doi: 10.1016/j.progpolymsci.2007.04.002.
64. Gomes, A. P., Mano, J. F., Queiroz, J. A. and Gouveia, I. C. (2015) 'Incorporation of antimicrobial peptides on functionalized cotton gauzes for medical applications', *Carbohydrate Polymers*. Elsevier Ltd., 127, pp. 451–461. doi: 10.1016/j.carbpol.2015.03.089.
65. Gonzales, M. F., Brooks, T., Pukatzki, S. U., & Provenzano, D. (2013). Rapid protocol for preparation of electrocompetent *Escherichia coli* and *Vibrio cholerae*. *Journal of Visualized Experiments*, 80, 6–11. <https://doi.org/10.3791/50684>
66. Guyot, N., Meudal, H., Trapp, S., Lochmann, S., Silvestre, A., Jousset, G., Labas, V., Reverdiau, P., Loth, K., Hervé, V., Aucagne, V., Delmas, A. F., Rehaut-Godbert, S., & Landon, C. (2020). Structure, function, and evolution of Gga-AvBD11, the archetype of the structural avian-double $\beta$ -defensin family. *Proceedings of the National Academy of Sciences of the United States of America*, 117(1), 337–345. <https://doi.org/10.1073/pnas.1912941117>
67. Hasegawa, M., Yang, K., Hashimoto, M., Park, J. H., Kim, Y. G., Fujimoto, Y., Nuñez, G., Fukase, K., & Inohara, N. (2006). Differential release and distribution of Nod1 and Nod2 immunostimulatory molecules among bacterial species and environments. *Journal of Biological Chemistry*, 281(39), 29054–29063. <https://doi.org/10.1074/jbc.M602638200>

68. Harwig, S. S. L., Swiderek, K. M., Kokryakov, V. N., Tan, L., Lee, T. D., Panyutich, E. A., Aleshina, G. M., Shamova, O. v., & Lehrer, R. I. (1994). Gallinacins: cysteine-rich antimicrobial peptides of chicken leukocytes. *FEBS Letters*, *342*(3), 281–285. [https://doi.org/10.1016/0014-5793\(94\)80517-2](https://doi.org/10.1016/0014-5793(94)80517-2)
69. Hargreaves, A. D., Swain, M. T., Hegarty, M. J., Logan, D. W. and Mulley, J. F. (2018) 'Restriction and Recruitment — Gene Duplication and the Origin and Evolution of Snake Venom Toxins', *6*(February), pp. 2088–2095. doi: 10.1093/gbe/evu166.
70. Hajji, M., Jellouli, K., Hmidet, N., Balti, R., Sellami-Kamoun, A. and Nasri, M. (2010) 'A highly thermostable antimicrobial peptide from *Aspergillus clavatus* ES1: Biochemical and molecular characterization', *Journal of Industrial Microbiology and Biotechnology*, *37*(8), pp. 805–813. doi: 10.1007/s10295-010-0725-6.
71. Hazlett, L. and Wu, M. (2011) Defensins in innate immunity. *Cell and Tissue Research*, **343**, 175–188.
72. Heise, P. J., Maxson, L. R., Dowling, H. G., & Hedges, S. B. (1995). Higher-level snake phylogeny inferred from mitochondrial DNA sequences of 12S rRNA and 16S rRNA genes. *Molecular Biology and Evolution*, *12*(2), 259–265. <https://doi.org/10.1093/oxfordjournals.molbev.a040202>
73. Hellgren, O. and Ekblom, R. (2010) 'Evolution of a cluster of innate immune genes (-defensins) along the ancestral lines of chicken and zebra finch', *Immunome Research*, *6*(1), pp. 1–15. doi: 10.1186/1745-7580-6-3.
74. Herbel, V., Schäfer, H. and Wink, M. (2015) 'Recombinant production of snakain-2 (an antimicrobial peptide from tomato) in *E. Coli* and analysis of its bioactivity', *Molecules*, *20*(8), pp. 14889–14901. doi: 10.3390/molecules200814889.
75. Hirsch T, Spielmann M, Zuhaili B, Fossum M, Metzsig M, Koehler T, Steinau HU, Yao F, Onderdonk AB, Steinstaesser L, Eriksson E (2009) Human beta-defensin-3 promotes wound healing in infected diabetic wounds, *J. Gene. Med*, **11**, 220-228.
76. Higgs, R., Lynn, D. J., Gaines, S., McMahon, J., Tierney, J., James, T., Lloyd, A. T., Mulcahy, G., & O'Farrelly, C. (2005). The synthetic form of a novel chicken  $\beta$ -defensin identified in silico is predominantly active against intestinal pathogens. *Immunogenetics*, *57*(1–2), 90–98. <https://doi.org/10.1007/s00251-005-0777-3>
77. Hughes A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings. Biological sciences*, *256*(1346), 119–124. <https://doi.org/10.1098/rspb.1994.0058>
78. Hughes, A. L. (1999) 'Evolutionary diversification of the mammalian defensins', *Cellular and Molecular Life Sciences*, *56*(1–2), pp. 94–103. doi: 10.1007/s000180050010.

79. Hultmark D, Steiner H, Rasmuson T, Boman HG (1980) Insect Immunity. Purification and properties of three inducible bactericidal proteins from the hemolymph of the immobilized pupae of *Hyalophora cecropia*. *Eur J Biochem* **106**, 7-16
80. Hurles, M. (2004) 'Gene Duplication: The Genomic Trade in Spare Parts', 2(7). doi: 10.1371/journal.pbio.0020206.
81. Inohara, N., Koseki, T., Lin, J., del Peso, L., Lucas, P. C., Chen, F. F., Ogura, Y., & Núñez, G. (2000). An induced proximity model for NF-kappa B activation in the Nod1/RICK and RIP signaling pathways. *The Journal of Biological Chemistry*, 275(36), 27823–27831. <https://doi.org/10.1074/jbc.M003415200>
82. Jalkanen, J., Huhtaniemi, I., & Poutanen, M. (2005). Discovery and characterization of new epididymis-specific beta-defensins in mice. *Biochimica et Biophysica Acta - Gene Structure and Expression*, 1730(1), 22–30. <https://doi.org/10.1016/j.bbaexp.2005.05.010>
83. Kaessmann, H. (2010) 'Origins, evolution, and phenotypic impact of new genes', *Cold Spring Harbor Laboratory Press*, 20(10), pp. 1313–1326. doi: 10.1101/gr.101386.109.
84. Kaplinsky, N. J., Gilbert, S. F., Cebra-Thomas, J., Lilleväli, K., Saare, M., Chang, E. Y., Edelman, H. E., Frick, M. A., Guan, Y., Hammond, R. M., Hampilos, N. H., Opoku, D. S. B., Sariahmed, K., Sherman, E. A., & Watson, R. (2013). The Embryonic Transcriptome of the Red-Eared Slider Turtle (*Trachemys scripta*). *PLoS ONE*, 8(6). <https://doi.org/10.1371/journal.pone.0066357>
85. Kapust, R. B., Toözseór, J., Copeland, T. D., & Waugh, D. S. (2002). The P1' specificity of tobacco etch virus protease. *Biochemical and Biophysical Research Communications*, 294(5), 949–955. [https://doi.org/10.1016/S0006-291X\(02\)00574-0](https://doi.org/10.1016/S0006-291X(02)00574-0)
86. Kawasaki, T. and Kawai, T. (2014) 'Toll-like receptor signaling pathways', *Frontiers in Immunology*, 5(SEP), pp. 1–8. doi: 10.3389/fimmu.2014.00461.
87. Kingshott, P., Wei, J., Bagge-Ravn, D., Gadegaard, N. and Gram, L. (2003) 'Covalent Attachment of Polyethylene glycol to Surfaces, Critical for Reducing Bacterial Adhesion', *Langmuir*. American Chemical Society, 19(17), pp. 6912–6921. doi: 10.1021/la034032m.
88. Klint, J. K., Senff, S., Saez, N. J., Seshadri, R., Lau, H. Y., Bende, N. S., Undheim, E. A. B., Rash, L. D., Mobli, M., & King, G. F. (2013). Production of Recombinant Disulfide-Rich Venom Peptides for Structural and Functional Analysis via Expression in the Periplasm of *E. coli*. *PLoS ONE*, 8(5). <https://doi.org/10.1371/journal.pone.0063865>
89. Kent, W. J. (2002) 'BLAT---The BLAST-Like Alignment Tool', *Genome Research*, 12(4), pp. 656–664. doi: 10.1101/gr.229202.
90. Korber B. (2000) HIV Signature and Sequence Variation Analysis. Computational Analysis of HIV Molecular Sequences, Chapter 4, pages 55-72. Allen G. Rodrigo and Gerald H. Learn, eds. Dordrecht, Netherlands: Kluwer Academic Publishers.

91. Kosakovsky Pond, S. L., Frost, S. D. W. and Muse, S. V. (2005) 'HyPhy: Hypothesis testing using phylogenies', *Bioinformatics*, 21(5), pp. 676–679. doi: 10.1093/bioinformatics/bti079.
92. Kosakovsky Pond, S. L. and Frost, S. D. W. (2005a) 'Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments', *Bioinformatics*, 21(10), pp. 2531–2533. doi: 10.1093/bioinformatics/bti320.
93. Kosakovsky Pond, S.L and Frost, S. D. W. (2005b) 'Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection', *Molecular Biology and Evolution*, 22(5), pp. 1208–1222, <https://doi.org/10.1093/molbev/msi105>
94. Kosikowska, P. and Lesner, A. (2016) 'Antimicrobial peptides (AMPs) as drug candidates: a patent review (2003-2015).', *Expert opinion on therapeutic patents*, 26(6), pp. 689–702. doi: 10.1080/13543776.2016.1176149.
95. Lai, Y. and Gallo, R.L. (2009) AMPed up immunity: how antimicrobial peptides have multiple roles in immune defence. *Trends in Immunology*, **30**, 131–141.
96. Lakshminarayanan, R., Vivekanandan, S., Samy, R. P., Banerjee, Y., Chi-Jin, E. O., Kay, W. T., Jois, S. D. S., Kini, R. M., & Valiyaveetil, S. (2008). Structure, self-assembly, and dual role of a  $\beta$ -defensin-like peptide from the Chinese soft-shelled turtle eggshell matrix. *Journal of the American Chemical Society*, 130(14), 4660–4668. <https://doi.org/10.1021/ja075659k>
97. Lan, C.C.E., Wu, C.S., Huang, S.M., Kuo, H.Y., Wu, I.H., Wen, C.H., Chai, C.Y., Fang, A.H., and Chen, G.S. (2011) High-glucose environment inhibits p38MAPK signaling and reduces human beta-3 expression in keratinocytes. *Molecular Medicine*, **17**, 771–779. <sup>[L]</sup><sub>[SEP]</sub>
98. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947-2948.
99. Leulier, F. and Lemaitre, B. (2008) 'Toll-like receptors - Taking an evolutionary approach', *Nature Reviews Genetics*, 9(3), pp. 165–178. doi: 10.1038/nrg2303.
100. Lee, M. S. and Kim, Y. J. (2007) 'Signaling pathways downstream of pattern-recognition receptors and their cross talk', *Annual Review of Biochemistry*, 76(II), pp. 447–480. doi: 10.1146/annurev.biochem.76.060605.122847.
101. Li, X. and Leong, S. S. J. (2011) 'A chromatography-focused bioprocess that eliminates soluble aggregation for bioactive production of a new antimicrobial peptide candidate', *Journal of Chromatography A*. Elsevier B.V., 1218(23), pp. 3654–3659. doi: 10.1016/j.chroma.2011.04.017.
102. Li, C. L., Xu, T. T., Chen, R. B., Huang, X. X., Zhao, Y. C., Bao, Y. Y., Zhao, W. D. and Zheng, Z. Y. (2013) 'Cloning, expression and characterization of antimicrobial porcine  $\beta$  defensin 1 in

Escherichia coli', *Protein Expression and Purification*, 88(1), pp. 47–53. doi: 10.1016/j.pep.2012.11.015.

103. Li, L., Brunk, B.P., Kissinger, J.C., Pape, D., Tang, K., Cole, R.H., Martin, J., Wylie, T., Dante, M., Fogarty, S.J. and Howe, D.K., 2003. Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. *Genome Research*, 13(3), pp.443-454.

104. Li, Y. (2011). Recombinant production of antimicrobial peptides in Escherichia coli: A review. *Protein Expression and Purification*, 80(2), 260–267. <https://doi.org/10.1016/j.pep.2011.08.001>

105. Li, Y. (2013). Recombinant production of crab antimicrobial protein scygonadin expressed as thioredoxin and SUMO fusions in escherichia coli. *Applied Biochemistry and Biotechnology*, 169(6), 1847–1857. <https://doi.org/10.1007/s12010-013-0102-9>

106. Lin, C. H., Pan, Y. C., Liu, F. W., & Chen, C. Y. (2017). Prokaryotic expression and action mechanism of antimicrobial LsGRP1C recombinant protein containing a fusion partner of small ubiquitin-like modifier. *Applied Microbiology and Biotechnology*, 101(22), 8129–8138. <https://doi.org/10.1007/s00253-017-8530-z>

107. Luan, C., Zhang, H. W., Song, D. G., Xie, Y. G., Feng, J., & Wang, Y. Z. (2014). Expressing antimicrobial peptide cathelicidin-BF in Bacillus subtilis using SUMO technology. *Applied Microbiology and Biotechnology*, 98(8), 3651–3658. <https://doi.org/10.1007/s00253-013-5246-6>

108. Lynn, D. J., Higgs, R., Gaines, S., Tierney, J., James, T., Lloyd, A. T., Fares, M. A., Mulcahy, G., & O'Farrelly, C. (2004). Bioinformatic discovery and initial characterisation of nine novel antimicrobial peptide genes in the chicken. *Immunogenetics*, 56(3), 170–177. <https://doi.org/10.1007/s00251-004-0675-0>

109. Lynn, D. J., Higgs, R., Lloyd, A. T., O'Farrelly, C., Hervé-Grépinet, V., Nys, Y., Brinkman, F. S. L., Yu, P. L., Soulier, A., Kaiser, P., Zhang, G., & Lehrer, R. I. (2007). Avian beta-defensin nomenclature: A community proposed update. *Immunology Letters*, 110(1), 86–89. <https://doi.org/10.1016/j.imlet.2007.03.007>

110. Mable B. 2004. "Why polyploidy is rarer in animals than in plants": myths and mechanisms. *Biol J Linn Soc.* 82:453–466.

111. Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*. 2019 Jul;47(W1):W636-W641. DOI: 10.1093/nar/gkz268.

112. Malmsten, M. (2014) 'Antimicrobial peptides', *Upsala Journal of Medical Sciences*, 119(2), pp. 199–204. doi: 10.3109/03009734.2014.899278.

113. Matsushima, N., Tanaka, T., Enkhbayar, P., Mikami, T., Taga, M., Yamada, K., & Kuroki, Y. (2007). Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-

like receptors. *BMC Genomics*, 8, 1–20. <https://doi.org/10.1186/1471-2164-8-124>

114. Maxwell, A. I., Morrison, G. M. and Dorin, J. R. (2003) 'Rapid sequence divergence in mammalian  $\beta$ -defensins by adaptive evolution', *Molecular Immunology*, 40(7), pp. 413–421. doi: 10.1016/S0161-5890(03)00160-3.

115. McDermott, A. M. (2009) 'The role of antimicrobial peptides at the ocular surface', *Ophthalmic Research*, 41(2), pp. 60–75. doi: 10.1159/000187622.

116. Medzhitov, R., Preston-Hurlburt, P. and Janeway, C. A. (1997) 'A human homologue of the *Drosophila* toll protein signals activation of adaptive immunity', *Nature*, 388(6640), pp. 394–397. doi: 10.1038/41131

117. Merchant, M., Morkotinis, V., Hale, A., White, M., & Moran, C. (2017). Crocodylian nuclear factor kappa B. *Comparative Biochemistry and Physiology Part - B: Biochemistry and Molecular Biology*, 213(July), 28–34. <https://doi.org/10.1016/j.cbpb.2017.07.009>

118. Michaelson, D., Rayner, J., Couto, M., & Ganz, T. (1992). Cationic defensins arise from charge-neutralized propeptides: a mechanism for avoiding leukocyte autotoxicity? *Journal of Leukocyte Biology*, 51(6), 634–639. <https://doi.org/10.1002/jlb.51.6.634>

119. Mount DM. (2004). *Bioinformatics: Sequence and Genome Analysis* (2nd ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. ISBN 0-87969-608-7.

120. Morrison, G. M., Semple, C. A. M., Kilanowski, F. M., Hill, R. E., & Dorin, J. R. (2003). Signal sequence conservation and mature peptide divergence within subgroups of the murine  $\beta$ -defensin gene family. *Molecular Biology and Evolution*, 20(3), 460–470. <https://doi.org/10.1093/molbev/msg060>

121. Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., Scheffler, K. (2013) 'FUBAR: A Fast, Unconstrained Bayesian Approximation for Inferring Selection', *Molecular Biology and Evolution*, 30(5), pp 1196–1205, <https://doi.org/10.1093/molbev/mst030>

122. Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., Kosakovsky Pond, S. L. (2012) 'Detecting Individual Sites Subject to Episodic Diversifying Selection.' *PLoS Genetics*, 8(7): e1002764. <https://doi.org/10.1371/journal.pgen.1002764>

123. Müller, H., Salzig, D. and Czermak, P. (2015) 'Considerations for the process development of insect-derived antimicrobial peptide production', *Biotechnology Progress*, 31(1), pp. 1–11. doi: 10.1002/btpr.2002.

124. Mygind, P. H., Fischer, R. L., Schnorr, K. M., Hansen, M. T., Sönksen, C. P., Ludvigsen, S., Raventós, D., Buskov, S., Christensen, B., De Maria, L., Taboureau, O., Yaver, D., Elvig-Jørgensen, S. G., Sørensen, M. V., Christensen, B. E., Kjaerulff, S., Frimodt-Møller, N., Lehrer, R. I., Zasloff, M. and Kristensen, H.-H. (2005) 'Plectasin is a peptide antibiotic with therapeutic potential from a saprophytic fungus.', *Nature*, 437(7061), pp. 975–80. doi:



10.1038/nature04051.

125. Nei, M., and A. L. Hughes (1992) 'Balanced polymorphism and evolution by the birth-and-death process in the MHC loci' Pp. 27–38 in K. Tsuji, M. Aizawa, and T. Sasazuki, eds. 11th Histocompatibility Workshop and Conference. Oxford University Press, Oxford.

126. Nei, M. and Gojobori, T. (1986) 'Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions', *Molecular Biology and Evolution*, 3(5), pp. 418–426. doi: 10.1093/oxfordjournals.molbev.a040410.

127. Nei, M., Gu, X. and Sitnikova, T. (1997) 'Evolution by the birth-and-death process in multigene families of the vertebrate immune system', *Proceedings of the National Academy of Sciences of the United States of America*, 94(15), pp. 7799–7806. doi: 10.1073/pnas.94.15.7799.

128. Nigro, G., Fazio, L. L., Martino, M. C., Rossi, G., Tattoli, I., Liparoti, V., de Castro, C., Molinaro, A., Philpott, D. J., & Bernardini, M. L. (2008). Muramylpeptide shedding modulates cell sensing of *Shigella flexneri*. *Cellular Microbiology*, 10(3), 682–695. <https://doi.org/10.1111/j.1462-5822.2007.01075.x>

129. Noé, L. and Kucherov, G. (2005) 'YASS: Enhancing the sensitivity of DNA similarity search', *Nucleic Acids Research*, 33(SUPPL. 2), pp. 540–543. doi: 10.1093/nar/gki478.

130. Ohno S. 1970. Evolution by gene duplication. Springer Verlag, Berlin.

131. Ohno S. 1972. So much "junk" DNA in our genome. Brookhaven Symp Biol 23: 366–370.

132. Oppenheim, J. J. and Yang, D. (2005) 'Alarmins: Chemotactic activators of immune responses', *Current Opinion in Immunology*, 17(4 SPEC. ISS.), pp. 359–365. doi: 10.1016/j.coi.2005.06.002.

133. Panteleev, P. V. and Ovchinnikova, T. V. (2017) 'Improved strategy for recombinant production and purification of antimicrobial peptide tachyplexin I and its analogs with high cell selectivity', *Biotechnology and Applied Biochemistry*, 64(1), pp. 35–42. doi: 10.1002/bab.1456.

134. Pasupuleti M, Schmidtchen A, Malmsten M. Antimicrobial peptides: key components of the innate immune system. *Crit Rev Biotechnol*. 2012;32:143–71.

135. Patil, A. A., Cai, Y., Sang, Y., Blecha, F. and Zhang, G. (2005). Cross-species analysis of the mammalian beta-defensin gene family: presence of syntenic gene clusters and preferential expression in the male reproductive tract. *Physiol Genomics* 23,5-17.

136. Patil, A., Hughes, A. L. and Zhang, G. (2005) 'Rapid evolution and diversification of mammalian  $\alpha$ -defensins as revealed by comparative analysis of rodent and primate genes', *Physiological Genomics*, 20(38), pp. 1–11. doi: 10.1152/physiolgenomics.00150.2004.

137. Patterson-Delafield J, Szklarek D, Martinez RJ, Lehrer RI (1981) Microbial cationic proteins of rabbit alveolar macrophages: amino acid composition and functional attributes. *Infect Immun* **31**, 723-731
138. Perez Espitia, P. J., de Fátima Ferreira Soares, N., dos Reis Coimbra, J. S., de Andrade, N. J., Souza Cruz, R. and Alves Medeiros, E. A. (2012) 'Bioactive Peptides: Synthesis, Properties, and Applications in the Packaging and Preservation of Food', *Comprehensive Reviews in Food Science and Food Safety*, 11(2), pp. 187–204. doi: 10.1111/j.1541-4337.2011.00179.x.
139. Phoenix, D. A., Dennison, S. R. and Harris, F. (2013) Antimicrobial Peptides: Their History, Evolution, and Functional Promiscuity, in Antimicrobial Peptides, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany. doi: 10.1002/9783527652853.ch1
140. de Queiroz, A., & Groen, R. R. (2001). The Inconsistent and Inefficient Constricting Behavior of Colorado Western Terrestrial Garter Snakes, *Thamnophis elegans*. *Journal of Herpetology*, 35(3), 450–460. <https://doi.org/10.2307/1565963>
141. Radhakrishnan, Y., Fares, M. A., French, F. S., & Hall, S. H. (2007). Comparative genomic analysis of a mammalian  $\beta$ -defensin gene cluster. *Physiological Genomics*, 30(3), 213–222. <https://doi.org/10.1152/physiolgenomics.00263.2006>
142. Reese, M. G. (2001) 'Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome', *Computers and Chemistry*, 26(1), pp. 51–56. doi: 10.1016/S0097-8485(01)00099-7.
143. Reese MG, Eeckman, FH, Kulp, D, Haussler, D, 1997. "Improved Splice Site Detection in Genie". *J Comp Biol* **4(3)**, 311-23.
144. Rios, Francesca M, and Zimmerman, Laura M (Oct 2015) Immunology of Reptiles. In: eLS. John Wiley & Sons Ltd, Chichester. <http://www.els.net> [doi: 10.1002/9780470015902.a0026260]
145. Ringstad L, Schmidtchen A, Malmsten M. Effect of peptide length on the interaction between consensus peptides and DOPC/DOPA bilayers. *Langmuir*. 2006;22:5042–50.
146. Ringstad L, Andersson Nordahl E, Schmidtchen A, Malmsten M. Composition effect on peptide interaction with lipids and bacteria: variants of C3a peptide CNY21. *Biophys J*. 2007;92:87–98.
147. Salwiczek, M., Qu, Y., Gardiner, J., Strugnell, R. A., Lithgow, T., McLean, K. M. and Thissen, H. (2014) 'Emerging rules for effective antimicrobial coatings', *Trends in Biotechnology*. Elsevier Ltd, 32(2), pp. 82–90. doi: 10.1016/j.tibtech.2013.09.008.

148. Santana, F. L., Estrada, K., Ortiz, E., & Corzo, G. (2021). Reptilian  $\beta$ -defensins: Expanding the repertoire of known crocodylian peptides. *Peptides*, 136(September 2020), 170473. <https://doi.org/10.1016/j.peptides.2020.170473>
149. Schneider, J. G. (1801). "Porosus". *Historiae Amphibiorum naturalis et literariae Fasciculus Secundus continens Crocodilos, Scincos, Chamaesauras, Boas, Pseudoboas, Elapes, Angues, Amphisbaenas et Caecilias*. Jenae: Wesselhoeft. pp. 159–160.
150. Schneider, M., Zimmermann, A. G., Roberts, R. A., Zhang, L., Karen, V., Rahman, A. H., Conti, B. J., Eitas, T. K., & Koller, B. H. (2013). *The innate immune sensor NLRC3 attenuates Toll-like receptor signaling via modification of the signaling adaptor TRAF6 and transcription factor NF- $\kappa$ B*. 13(9), 823–831. <https://doi.org/10.1038/ni.2378>.The
151. Schutte, B. C., Mitros, J. P., Bartlett, J. A., Walters, J. D., Jia, H. P., Welsh, M. J., Casavant, T. L., & McCray, P. B. (2002). Discovery of five conserved  $\beta$ -defensin gene clusters using a computational search strategy. *Proceedings of the National Academy of Sciences of the United States of America*, 99(4), 2129–2133. <https://doi.org/10.1073/pnas.042692699>
152. Semple, C. A. M., Rolfe, M. and Dorin, J. R. (2003) 'Duplication and selection in the evolution of primate beta-defensin genes.', *Genome biology*, 4(5), p. R31. doi: 10.1186/gb-2003-4-5-r31.
153. Semple, C. A. M., Maxwell, A., Gautier, P., Kilanowski, F. M., Eastwood, H., Barran, P. E., & Dorin, J. R. (2005). The complexity of selection at the major primate  $\beta$ -defensin locus. *BMC Evolutionary Biology*, 5, 1–14. <https://doi.org/10.1186/1471-2148-5-32>
154. Semple, C.A., Gautier, P., Taylor, K., and Dorin, J.R. (2006) The changing of the guard: molecular diversity and rapid evolution of beta-defensins. *Molecular Diversity*, **10**, 575–584.
155. Semple, C. A. M. *et al.* (2006) ' $\beta$ -Defensin evolution: Selection complexity and clues for residues of functional importance', *Biochemical Society Transactions*, 34(2), pp. 257–262. doi: 10.1042/BST20060257.
156. Semple, F. and Dorin, J. R. (2012) ' $\beta$ -Defensins: Multifunctional modulators of infection, inflammation and more?', *Journal of Innate Immunity*, 4(4), pp. 337–348. doi: 10.1159/000336619.
157. Selsted, M. E. and Ouellette, A. J. (2005) 'Mammalian defensins in the antimicrobial immune response', *Nature Immunology*, 6(6), pp. 551–557. doi: 10.1038/ni1206.
158. Shedlock, A. M., Botka, C. W., Zhao, S., Shetty, J., Zhang, T., Liu, J. S., Deschavanne, P. J. and Edwards, S. V. (2007) 'Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome', *Proceedings of the National Academy of Sciences*, 104(8), pp. 2767–2772. doi: 10.1073/pnas.0606204104.

159. Sigurdadottir T, Andersson P, Davoudi M, Malmsten M, Schmidtchen A, Bodelsson M. In silico identification and biological evaluation of antimicrobial peptides based on human cathelicidin LL-37. *Antimicrob Agents Chemother*. 2006;50:2983–9.
160. Skouri-Gargouri, H. and Gargouri, A. (2008) 'First isolation of a novel thermostable antifungal peptide secreted by *Aspergillus clavatus*', *Peptides*, 29(11), pp. 1871–1877. doi: 10.1016/j.peptides.2008.07.005.
161. Singha, P., Locklin, J. and Handa, H. (2017) 'A review of the recent advances in antimicrobial coatings for urinary catheters', *Acta Biomaterialia*. Acta Materialia Inc., 50, pp. 20–40. doi: 10.1016/j.actbio.2016.11.070.
162. Skouri-Gargouri, H. and Gargouri, A. (2008) 'First isolation of a novel thermostable antifungal peptide secreted by *Aspergillus clavatus*', *Peptides*, 29(11), pp. 1871–1877. doi: 10.1016/j.peptides.2008.07.005.
163. Soman, S. S., Arathy, D. S. and Sreekumar, E. (2009) 'Discovery of *Anas platyrhynchos* avian beta-defensin 2 (Apl\_AvBD2) with antibacterial and chemotactic functions', *Molecular Immunology*, 46(10), pp. 2029–2038. doi: 10.1016/j.molimm.2009.03.003.
164. Solovyev, V., Kosarev, P., Seledsov, I., & Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biology*, 7 Suppl 1(Suppl 1), 1–12. <https://doi.org/10.1186/gb-2006-7-s1-s10>
165. Sørensen, H. P. and Mortensen, K. K. (2005) 'Advanced genetic strategies for recombinant protein expression in *Escherichia coli*', *Journal of Biotechnology*, 115(2), pp. 113–128. doi: <https://doi.org/10.1016/j.jbiotec.2004.08.004>.
166. Sorensen, O.E., Cowland, J.B., Theilgaard-Monch, K., Liu, L.D., Ganz, T., and Borregaard, N. (2003) Wound healing and expression of antimicrobial peptides/polypeptides in human keratinocytes, a consequence of common growth factors. *Journal of Immunology*, **170**, 5583–5589.
167. Stegemann, C., Kolobov, A., Leonova, Y. F., Knappe, D., Shamova, O., Ovchinnikova, T. v., Kokryakov, V. N., & Hoffmann, R. (2009). Isolation, purification and de novo sequencing of TBD-1, the first beta-defensin from leukocytes of reptiles. *Proteomics*, 9(5), 1364–1373. <https://doi.org/10.1002/pmic.200800569>
168. Steubesand, N., Kiehne, K., Brunke, G., Pahl, R., Reiss, K., Herzig, K. H., Schubert, S., Schreiber, S., Fölsch, U. R., Rosenstiel, P., & Arlt, A. (2009). The expression of the  $\beta$ -defensins hBD-2 and hBD-3 is differentially regulated by NF- $\kappa$ B and MAPK/AP-1 pathways in an in vitro model of *Candida esophagitis*. *BMC Immunology*, 10, 1–16. <https://doi.org/10.1186/1471-2172-10-36>
169. Sugiarto, H. and Yu, P. L. (2004) 'Avian antimicrobial peptides: The defence role of  $\beta$ -defensins', *Biochemical and Biophysical Research Communications*, 323(3), pp. 721–727. doi: 10.1016/j.bbrc.2004.08.162.

170. Suyama, M., Torrents, D. and Bork, P. (2006) 'PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments', *Nucleic Acids Research*, 34(WEB. SERV. ISS.), pp. 609–612. doi: 10.1093/nar/gkl315.
171. Tait, A. R. and Straus, S. K. (2011) 'Overexpression and purification of U24 from human herpesvirus type-6 in E. coli: Unconventional use of oxidizing environments with a maltose binding protein-hexahistidine dual tag to enhance membrane protein yield', *Microbial Cell Factories*, 10, pp. 1–12. doi: 10.1186/1475-2859-10-51
172. Tang, K. Y., Wang, X., Wan, Q. H., & Fang, S. G. (2018). A crucial role of paralogous  $\beta$ -defensin genes in the Chinese alligator innate immune system revealed by the first determination of a Crocodylia defensin cluster. *Developmental and Comparative Immunology*, 81, 193–203. <https://doi.org/10.1016/j.dci.2017.11.018>
173. Teng, L., Gao, B., and Zhang, S. (2012) The first chordate big defensin: identification, expression and bioactivity. *Fish & Shellfish Immunology*, 32, 572–577.
174. 2019 Antibacterial agents in clinical development: an analysis of the antibacterial clinical development pipeline. Geneva: World Health Organization; 2019. Licence: CC BY-NC-SA 3.0 IGO.
175. Thallinger, B., Argirova, M., Lesseva, M., Ludwig, R., Sygmund, C., Schlick, A., Nyanhongo, G. S. and Guebitz, G. M. (2014) 'Preventing microbial colonisation of catheters: Antimicrobial and antibiofilm activities of cellobiose dehydrogenase', *International Journal of Antimicrobial Agents*. Elsevier B.V., 44(5), pp. 402–408. doi: 10.1016/j.ijantimicag.2014.06.016.
176. Tsutsumi-ishii, Y. and Nagaoka, I. (2001) 'NF- $\kappa$ B-mediated transcriptional regulation of human  $\alpha$ -defensin-2 gene following lipopolysaccharide stimulation bacterial infection and proinflammatory cytokines.', *Journal of Leukocyte Biology*, 71(1), pp. 154–162.
177. Trifinopoulos, J., Nguyen, L. T., von Haeseler, A., & Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, 44(W1), W232–W235. <https://doi.org/10.1093/nar/gkw256>
178. Tu, J., Li, D., Li, Q., Zhang, L., Zhu, Q., Gaur, U., Fan, X., Xu, H., Yao, Y., Zhao, X., & Yang, M. (2015). Molecular evolutionary analysis of  $\beta$ -defensin peptides in vertebrates. *Evolutionary Bioinformatics*, 11, 105–114. <https://doi.org/10.4137/EBO.S25580>
179. Uehara, A., Yang, S., Fujimoto, Y., Fukase, K., Kusumoto, S., Shibata, K., Sugawara, S., & Takada, H. (2005). Muramyl dipeptide and diaminopimelic acid-containing desmuramyl peptides in combination with chemically synthesized Toll-like receptor agonists synergistically induced production of interleukin-8 in a NOD2- and NOD1-dependent manner, respectively, in human. *Cellular Microbiology*, 7(1), 53–61. <https://doi.org/10.1111/j.1462-5822.2004.00433.x>

180. Uehara, A., Fujimoto, Y., Kawasaki, A., Kusumoto, S., Fukase, K., & Takada, H. (2006). Meso -Diaminopimelic Acid and Meso -Lanthionine, Amino Acids Specific to Bacterial Peptidoglycans, Activate Human Epithelial Cells through NOD1. *The Journal of Immunology*, 177(3), 1796–1804. <https://doi.org/10.4049/jimmunol.177.3.1796>
181. van Dijk, A., Veldhuizen, E. J. A. and Haagsman, H. P. (2008) 'Avian defensins', *Veterinary Immunology and Immunopathology*, 124(1–2), pp. 1–18. doi: 10.1016/j.vetimm.2007.12.006.
182. van Hoek, M. L., Prickett, M. D., Settlege, R., Kang, L., Michalak, P., Vliet, K. A., & Bishop, B. M. (2019). The Komodo dragon (*Vranus komodoensis*) Genome and Identification of Innate Immunity Genes and Clusters. *BMC Genomics*, 1–18. [papers3://publication/uuid/BB1E342E-AE67-4FDE-AC8C-61D2406769E2](https://doi.org/10.1186/s12864-019-0699-2)
183. van Hoek, M. L. (2014) 'Antimicrobial peptides in reptiles', *Pharmaceuticals*, 7(6), pp. 723–753. doi: 10.3390/ph7060723.
184. Vriens, K., Cammue, B. P. A. and Thevissen, K. (2014) 'Antifungal plant defensins: Mechanisms of action and production', *Molecules*, 19(8), pp. 12280–12303. doi: 10.3390/molecules190812280.
185. Vu, T. T. T., Jeong, B., Yu, J., Koo, B.-K., Jo, S.-H., Robinson, R. C., & Choe, H. (2014). Soluble prokaryotic expression and purification of crostamine using an N-terminal maltose-binding protein tag. *Toxicon*, 92, 157–165. [https://doi.org/https://doi.org/10.1016/j.toxicon.2014.10.017](https://doi.org/10.1016/j.toxicon.2014.10.017)
186. Wan, Q. H., Pan, S. K., Hu, L., Zhu, Y., Xu, P. W., Xia, J. Q., Chen, H., He, G. Y., He, J., Ni, X. W., Hou, H. L., Liao, S. G., Yang, H. Q., Chen, Y., Gao, S. K., Ge, Y. F., Cao, C. C., Li, P. F., Fang, L. M., ... Fang, S. G. (2013). Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator. *Cell Research*, 23(9), 1091–1105. <https://doi.org/10.1038/cr.2013.104>
187. Whittington, C. M., Papenfuss, A. T., Bansal, P., Torres, A. M., Wong, E. S. W., Deakin, J. E., Graves, T., Alsop, A., Schatzkamer, K., Kremitzki, C., Ponting, C. P., Temple-Smith, P., Warren, W. C., Kuchel, P. W., & Belov, K. (2008). Defensins and the convergent evolution of platypus and reptile venom genes. *Genome Research*. <https://doi.org/10.1101/gr.7149808>
188. Webb, G. J. W., Manolis, S. C. and Brien, M. L. (2010) 'Saltwater Crocodile *Crocodylus porosus*', *Crocodyles.*, 3, pp. 99–113.
189. Wibowo, D. and Zhao, C. X. (2019) 'Recent achievements and perspectives for large-scale recombinant production of antimicrobial peptides', *Applied Microbiology and Biotechnology*. doi: 10.1007/s00253-018-9524-1.
190. Wilson C.L., Schmidt A.P., Pirila E, Valore E.V., Ferri N, Sorsa T, Ganz T Parks W.C. (2009) Differential Processing of {alpha}- and {beta}- Defensin Precursors by Matrix

Metalloproteinase-7 (MMP-7). *J. Biol. Chem.* **284**, 8301-8311.

191. Wlasiuk, G. and Nachman, M. W. (2010) 'Adaptation and Constraint at Toll-Like Receptors in Primates', *Molecular Biology and Evolution*, 27(9), pp. 2172–2186. doi: 10.1093/molbev/msq104.

192. Wei, X., Wu, R., Zhang, L., Ahmad, B., Si, D., & Zhang, R. (2018). Expression, purification, and characterization of a novel hybrid peptide with potent antibacterial activity. *Molecules*, 23(6). <https://doi.org/10.3390/molecules23061491>

193. Xiao, Y., Hughes, A. L., Ando, J., Matsuda, Y., Cheng, J. F., Skinner-Noble, D., & Zhang, G. (2004). A genome-wide screen identifies a single  $\beta$ -defensin gene cluster in the chicken: Implications for the origin and evolution of mammalian defensins. *BMC Genomics*, 5, 1–11. <https://doi.org/10.1186/1471-2164-5-56>

194. Xu, B. and Yang, Z. (2013) 'PamlX: A graphical user interface for PAML', *Molecular Biology and Evolution*, 30(12), pp. 2723–2724. doi: 10.1093/molbev/mst179.

195. Yadav, D. K., Yadav, N., Yadav, S., Haque, S., & Tuteja, N. (2016). An insight into fusion technology aiding efficient recombinant protein production for functional proteomics. *Archives of Biochemistry and Biophysics*, 612, 57–77. <https://doi.org/https://doi.org/10.1016/j.abb.2016.10.012>

196. Yang, D. and Oppenheim, J. J. (2003) 'Defensins', in Henry, H. L. and Norman, A. W. (eds) *Encyclopedia of Hormones*. New York: Academic Press, pp. 385–392. doi: <https://doi.org/10.1016/B0-12-341103-3/00063-2>.

197. Yang, Z. and Nielsen, R. (1998) 'Synonymous and nonsynonymous rate variation in nuclear genes of mammals', *Journal of Molecular Evolution*, 46(4), pp. 409–418. doi: 10.1007/PL00006320.

198. Yang, Z. and Nielsen, R. (2000) 'Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models', *Molecular Biology and Evolution*, 17(1), pp. 32–43. doi: 10.1093/oxfordjournals.molbev.a026236.

199. Yang Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8), 1586–1591. <https://doi.org/10.1093/molbev/msm088>

200. Yang, D., Chertov, O., Bykovskaia, N., Chen, Q., Buffo, M.J., Shogan, J., Anderson, M., Schroder, J.M., Wang, J.M., Howard, O.M.Z., and Oppenheim, J.J. (1999) Beta-defensins: linking innate and adaptive immunity through dendritic and T cell CCR6. *Science*, **286**, 525-528. <sup>[1]</sup><sub>[SEP]</sub>

201. Yang, D., Biragyn, A., Kwak, L.W., and Oppenheim, J.J. (2002) Mammalian defensins in immunity: more than just microbicidal. *Trends in Immunology*, **23**, 291–296.

202. Yoneyama, M. and Fujita, T. (2009) 'RNA recognition and signal transduction by RIG-I-like receptors', *Immunological Reviews*, 227(1), pp. 54–65. doi: 10.1111/j.1600-065X.2008.00727.x.
203. Yount, N. Y., Kupferwasser, D., Spisni, A., Dutz, S. M., Ramjan, Z. H., Sharma, S., Waring, A. J., & Yeaman, M. R. (2009). Selective reciprocity in antimicrobial activity versus cytotoxicity of hBD-2 and crotamine. *Proceedings of the National Academy of Sciences of the United States of America*, 106(35), 14972–14977. <https://doi.org/10.1073/pnas.0904465106>
204. Yu, H., Liu, X., Wang, H., Feng, L., Qiao, X., Cai, S., Shi, N. and Wang, Y. (2016) 'Identification, eukaryotic expression and structure & function characterizations of  $\beta$ -defensin like homologues from *Pelodiscus sinensis*', *Developmental & Comparative Immunology*. Elsevier Ltd, 68, pp. 108–117. doi: 10.1016/j.dci.2016.11.020.
205. Zasloff M. (2002). Antimicrobial peptides in health and disease. *The New England journal of medicine*, 347(15), 1199–1200. <https://doi.org/10.1056/NEJMe020106>
206. Zhang, J., Yang, Y., Teng, D., Tian, Z., Wang, S. and Wang, J. (2011) 'Expression of plectasin in *Pichia pastoris* and its characterization as a new antimicrobial peptide against *Staphylococcus* and *Streptococcus*', *Protein Expression and Purification*, 78(2), pp. 189–196. doi: 10.1016/j.pep.2011.04.014.
207. Zhang, J. (2003) 'Evolution by gene duplication: An update', *Trends in Ecology and Evolution*, 18(6), pp. 292–298. doi: 10.1016/S0169-5347(03)00033-8.
208. Zhao, C., Nguyen, T., Liu, L., Sacco, R. E., Brogden, K. A., & Lehrer, R. I. (2001). Gallinacin-3, an inducible epithelial  $\beta$ -defensin in the chicken. *Infection and Immunity*, 69(4), 2684–2691. <https://doi.org/10.1128/IAI.69.4.2684-2691.2001>
209. Zhou, Y., Liang, Q., Li, W., Gu, Y., Liao, X., Fang, W., & Li, X. (2016). Characterization and functional analysis of toll-like receptor 4 in Chinese soft-shelled turtle *Pelodiscus sinensis*. *Developmental and Comparative Immunology*, 63, 128–135. <https://doi.org/10.1016/j.dci.2016.05.023>
210. Zimmerman, L. M. (2020) 'The reptilian perspective on vertebrate immunity: 10 years of progress', *The Journal of experimental biology*, 223. doi: 10.1242/jeb.214171.
211. Zhu, S. (2008) Discovery of six families of fungal defensin-like peptides provides insights into origin and evolution of the CS $\alpha$  $\beta$  defensins. *Molecular Immunology*, 45, 828–838.
212. Zhu, S. and Gao, B. (2012) Evolutionary origin of  $\beta$ -defensins. *Developmental & Comparative Immunology*, doi: 10.1016/j.dci.2012.02.011.
213. Zou, J., Mercier, C., Koussounadis, A. and Secombes, C. (2007) 'Discovery of multiple beta-defensin like homologues in teleost fish', *Molecular Immunology*, 44(4), pp. 638–647. doi: 10.1016/j.molimm.2006.01.012.



## APPENDIX

### A1.1 *Podarcis muralis* exon positions

GENE	EXON 1		Length (bp)	EXON 2		Length (bp)	Orientation	Total Length (bp)	No of AA
	FROM	END		FROM	END				
PMBD1	42399	42456	58	44853	44986	134	+	192	64
PMBD2	53955	54012	58	51785	51918	134	-	192	64
PMBD3	64410	64461	52	62553	62707	155	-	207	69
PMBD4	88728	88785	58	86662	86837	176	-	234	78
PMBD5	94047	94104	58	95133	95281	149	+	207	69
PMBD6	99725	99782	58	101125	101240	116	+	174	58
PMBD7	110708	110765	58	113331	113455	125	+	183	61
PMBD8	132592	132649	58	131415	131533	119	-	177	59
PMBD9	144944	145001	58	145966	146084	119	+	177	59
PMBD10	160297	160354	58	159119	159237	119	-	177	59
PMBD11	172503	172560	58	171090	171208	119	-	177	59
PMBD12	194279	194336	58	190945	191066	122	-	180	60
PMBD13	214060	214117	58	212818	212939	122	-	180	60
PMBD14	306898	306955	58	297933	298054	122	-	180	60
PMBD15	327152	327209	58	324308	324426	119	-	177	59
PMBD16	351419	351476	58	352561	352679	119	+	177	59
PMBD17	358609	358666	58	359754	359872	119	+	177	59
PMBD18	419434	419491	58	415785	415903	119	-	177	59
PMBD19	427807	427864	58	428944	429065	122	+	180	60
PMBD20	451788	451845	58	448138	448259	122	-	180	60
PMBD21	462469	462526	58	464344	464462	119	+	177	59
PMBD22	485872	485929	58	481920	482038	119	-	177	59
PMBD23	494284	494341	58	496166	496284	119	+	177	59
PMBD24	508431	508488	58	508890	509005	116	+	174	58
PMBD25	585494	585638	145	586807	586964	158	+	303	101
PMBD26	632455	632584	130	633784	633953	170	+	300	100
PMBD27	636366	636489	124	639274	639404	131	+	255	85
PMBD28	645464	645593	130	646791	646960	170	+	300	100
PMBD29	649377	649500	124	650804	650973	170	+	294	98
PMBD30	658583	658688	106	660407	660537	131	+	237	79
PMBD31	664657	664813	157	667993	668120	128	+	285	95

PMBD32	676911	677077	167	673773	673917	145	-	312	104
PMBD33	679300	679465	166	681173	681318	146	+	312	104
PMBD34	686273	686342	70	684820	684947	128	-	198	66
PMBD35	702757	702814	58	698463	698581	119	-	177	59
PMBD36	714997	715054	58	715989	716167	179	+	237	79
PMBD37	730221	730278	58	731232	731410	179	+	237	79
PMBD38	775952	776009	58	771733	771851	119	-	177	59
PMBD39	795793	795850	58	796810	796988	179	+	237	79
PMBD40	806402	806459	58	807401	807579	179	+	237	79
PMBD41	834661	834718	58	836056	836234	179	+	237	79
PMBD42	851691	851748	58	852721	852899	179	+	237	79
PMBD43	889008	889065	58	890475	890617	143	+	201	67
PMBD44	936333	936390	58	935250	935374	125	+	183	61
PMBD45	947649	947712	64	947143	947270	128	-	192	64
PMBD46	959102	959147	46	958087	958211	125	-	171	57
PMBD47	967490	967535	46	966433	966557	125	-	171	57
PMBD48	973147	973192	46	972068	972192	125	-	171	57
PMBD49	980561	980606	46	979600	979730	131	-	177	59
PMBD50	994701	994758	58	996579	996715	137	+	195	65
PMBD51	1004349	1004406	58	1006468	1006598	131	+	189	63
PMBD52	1030639	1030696	58	1027575	1027708	134	-	192	64
PMBD53	1045206	1045263	58	1041810	1041940	131	-	189	63
PMBD54	1054499	1054556	58	1051482	1051612	131	-	189	63
PMBD55	1070651	1070708	58	1071973	1072103	131	+	189	63
PMBD56	1342193	1342250	58	1344072	1344211	140	+	198	66
PMBD57	1353637	1353694	58	1352265	1352383	119	-	177	59
PMBD58	1363436	1363493	58	1362273	1362412	140	-	198	66
PMBD59	1375909	1375966	58	1377805	1377944	140	+	198	66
PMBD60	1390049	1390106	58	1388661	1388779	119	-	177	59
PMBD61	1395172	1395229	58	1396253	1396392	140	+	198	66
PMBD62	1451335	1451392	58	1452357	1452517	161	+	219	73
PMBD63	1467784	1467844	61	1466336	1466484	149	-	210	70
PMBD64	1475628	1475688	61	1474717	1474850	134	-	195	65
PMBD65	1484683	1484743	61	1483243	1483379	137	-	198	66
PMBD66	1596979	1597036	58	1601787	1601926	140	+	198	66

PMBD67	1642015	1642078	64	1650161	1650279	119	+	183	61
PMBD68	1680986	1681043	58	1678182	1678303	122	-	180	60
PMBD69	1690082	1690145	64	1687762	1687895	134	-	198	66
PMBD70	1695122	1695185	64	1693202	1693332	131	-	195	65
PMBD71	1714111	1714174	64	1717335	1717474	140	+	204	68
PMBD72	1747881	1747944	64	1750159	1750403	245	+	309	103
PMBD73	1842457	1842508	52	1841663	1841808	146	-	198	66
PMBD74	1863091	1863142	52	1861312	1861457	146	-	198	66
PMBD75	1881911	1881968	58	1878993	1879111	119	-	177	59
PMBD76	1894969	1895026	58	1893017	1893141	125	-	183	61
PMBD77	1904753	1904810	58	1905955	1906079	125	+	183	61
PMBD78	1910931	1910997	67	1911707	1911849	143	+	210	70
PMBD79	1957855	1957912	58	1960528	1960655	128	+	186	62
PMBD80	1967932	1967989	58	1972118	1972248	131	+	189	63

*Positions coordinates of exons. Last codon of CTSB is a marker for the start of the beta-defensin cluster. Orientation of the gene and length of exons and full length of coding sequence.*

A1.2 *Podarcis muralis* Beta-defensin physical properties

	pI	Net Charge	Mr		pI	Net Charge	Mr
PMBD1	9.69	8	5112	PMBD41	5.12	-2	6569
PMBD2	9.69	8	5112	PMBD42	4.46	-6	6361
PMBD3	7.04	0	5255	PMBD43	6.7	0	4679
PMBD4	8.92	4	6341	PMBD44	4.48	-4	4061
PMBD5	7.79	1	5236	PMBD45	8.3	2	4672
PMBD6	7.76	1	3951	PMBD46	6.7	0	4360
PMBD7	8.3	2	4269	PMBD47	6.25	-1	4166
PMBD8	6.7	0	4149	PMBD48	5.46	-1	4327
PMBD9	4.68	-1	4155	PMBD49	4.25	-3	4499
PMBD10	3.92	-3	4009	PMBD50	8.33	2	4795
PMBD11	4.41	-2	4095	PMBD51	7.75	1	4560
PMBD12	6.14	0	4227	PMBD52	10.27	12	4905
PMBD13	6.14	0	4102	PMBD53	9.04	5	4803
PMBD14	8.96	4	4335	PMBD54	9.04	5	4803
PMBD15	8.64	3	4168	PMBD55	8.32	2	4644
PMBD16	6.1	0	4029	PMBD56	4.44	-3	4664
PMBD17	6.1	0	4029	PMBD57	4.65	-2	3696
PMBD18	7.78	1	3953	PMBD58	4.14	-4	4724
PMBD19	4.14	-2	4016	PMBD59	4.44	-3	4664
PMBD20	5.5	-1	3984	PMBD60	7.77	1	3625
PMBD21	8.35	2	3974	PMBD61	4.14	-4	4693
PMBD22	7.78	1	3825	PMBD62	3.71	-6	4962
PMBD23	8.35	2	3846	PMBD63	8.68	3	5300
PMBD24	8.32	2	3774	PMBD64	9.55	7	5101

PMBD25	5.09	-3	9441	PMBD65	8.6	3	5030
PMBD26	5.01	-3	9136	PMBD66	8.62	3	5134
PMBD27	7.75	1	7491	PMBD67	8.65	3	4361
PMBD28	5.01	-3	9136	PMBD68	10.24	10	4348
PMBD29	5.02	-3	8855	PMBD69	5.71	-1	5086
PMBD30	5.05	-1	6458	PMBD70	6	-1	4899
PMBD31	4.49	-6	8406	PMBD71	5.27	-1	5066
PMBD32	4.51	-8	9657	PMBD72	8.95	8	8835
PMBD33	4.73	-6	9459	PMBD73	7.04	0	4640
PMBD34	10.47	10	4966	PMBD74	6.42	0	5071
PMBD35	6.97	0	4066	PMBD75	8.92	4	4365
PMBD36	5.07	-3	6282	PMBD76	9.34	6	4364
PMBD37	5.45	-2	6223	PMBD77	9.34	6	4364
PMBD38	7.78	1	4166	PMBD78	7.57	1	5065
PMBD39	5.07	-3	6324	PMBD79	9.15	5	4645
PMBD40	4.67	-5	6433	PMBD80	8.95	4	4603

Physical properties of the *Podarcis muralis* mature peptide beta-defensins. All properties were achieved by using the protparam program on the ExPASy Server (Gasteiger, E. *et al.* 2005)

A1.3 *Podarcis muralis* Beta-defensin sequences – signal and pro-domain/mature peptides

	Signal Peptide	Pro-Mature Peptide	Probability
PMBD1	MRFRNLLIVAILGAFVLVSLGAG	LNLFARRCRRRAGKCRGNRCFYNEIEISTCYHTKIKCCREKD	0.4837
PMBD2	MRFRNLLIVAILGAFVLVSLGAG	LNLFARRCRRRAGKCRGNRCFYNEIEISTCYHTKIKCCREKD	0.4837
PMBD3	MPSLFPVAFLLCTLTPGSHA	RDTLKCHEDKGTCHHTLCPAQKIEKGSCYSGLCCVGLVHRITEL	0.7422
PMBD4	MFYLSLAVQVVLVFSTIAA	HRQGEVVKNLAQVVICNEGRGYCLDVQSRCPSGLVFNNNCPNKTMNKCCTPFAGRGV	0.5652
PMBD5	MRLLYSFAVVVLAFAVAAGHA	HMEGRLRPCNGGRGYCLDIQFQCPSGLQFINNTCPNPTMFSCCTPVQ	0.4065
PMBD6	MRFLHLSFALVFILFHVAG	QPSCGELGGYCQVPLTLNCPYGNIPANCGFNGNCCSK	0.8058
PMBD7	MRFLYSFALVFIFFHVAAAG	QPAKSCEELGGYCQVPLTLKCPYGYIPAMCGINGNCCSK	0.4273
PMBD8	MRFLYSFTLVFILFHVAG	QLFCETFGGSCHFPATTNCTYGVKVPWVSCGDNGICSSK	0.7444
PMBD9	MRFLYSFTLVFILFHVAG	KLFCETFGGSQCQFPATTNCTYGEIPWVSCGDNGICSSK	0.6745
PMBD10	MRFLYLTALVFIHLVAG	QLFCEILGGTCQFPATTDCTYGEITGASCGDNGICCMGK	0.8908
PMBD11	MRFLYLSLTLVSILFHVAG	QRLCELFGGSCQFPATTNCTYEEITGASCGDNGICSSK	0.8328
PMBD12	MRFLYSFVLVFIFFHVAA	DEPPLCSLSGGFCQAPITANCPSEITFVPCGPNARCCRSK	0.9558
PMBD13	MRFLYSFTLVFILFHVAG	QDLLCELLGGTCQFPATRNCHANGEIRGVSCGSNGICCLGK	0.8426
PMBD14	MRFLYSFALVFIFFHVFAG	SSPLCRVLGGYCQAPRTVNCPFGENTLAFCGPNARCCRSK	0.4456
PMBD15	MRFLYSFALVFIFFHVAG	QLTSCKEQGGYCQVPLTLKCPYGNIPANCGFNGNCCSK	0.5636
PMBD16	MRFLYLIFTLVFIHLFHVAG	QRLCEILGGSCQFPATTNCTYGEIPVSCGTNGICSSSK	0.8561
PMBD17	MRFLYLIFTLVFIHLFHVAG	QRLCEILGGSCQFPATTNCTYGEIPVSCGTNGICSSSK	0.8561
PMBD18	MRFLYSFTLVIIIFHVAG	QFICELAGGNCHYSATTKCHANGEIKGISCGSNGICCNK	0.7524
PMBD19	MRFLYLIFTLVILFHVAG	QDLLCELLGGTCQFPATINCANGEIIGVSCGSNGICCLGK	0.8246
PMBD20	MRFLYSFTLVIIIFHVAG	QFICELAGGTCHYSATTNCENGEIKGISCGSNGICCNK	0.7439
PMBD21	MRFLYLTALVFIHLVAV	QLGCERLGGTCQFPATRNCHANGEIRGVSCGSNGICCLGK	0.2034
PMBD22	MRFLYSFTLVIIIFHVAGQ	FICELAGGNCHYSATTKCHANGEIKGISCGSNGICCNK	0.7524
PMBD23	MRFLYLTALVFIHLVAEQ	LGGERLGGTCQFPATRNCHANGEIRGVSCGSNGICCLGK	0.5774
PMBD24	MKFFHIFAMVLLFQVFTGVHL	IPCDEMGGYCVIKPALCEPQIRGYCGPNRKCCKA	0.6936
PMBD25	MKFVCLFFALVFLCSAQA	DEADLAKMEKQEGENLKDQLQEEDPAGDDQDSGPKASPRLAVVGCYGNRGYCLPRGYRCHNGLKWEKEYNNCPYRNVLCCVR	0.9109
PMBD26	MKMLYVLTVAFLVFQVWS	NPKPPSKVEGEAKEFLESRAYKDDGPGLNPEPKDSSRFQVVLCSNDNDGYCLPRDFQCHNGLAFKEPWNDPFSVLLKCCVR	0.917
PMBD27	MKMLYVLFVAVFLVYQVQA	NPKPPEDEAKEPLDPEMKLKEEDVPRPRNAMVCNAMGGNCRSRCNDNEKSIGKCFASRYCCVRFQ	0.9686
PMBD28	MKMLYVLTVAFLVFQVWS	NPKPPSKVEGEAKEFLESRAYKDDGPGLNPEPKDSSRFQVVLCSNDNDGYCLPRDFQCHNGLAFKEPWNDPFSVLLKCCVR	0.917
PMBD29	MKMLYVLFVAVFLVYQVQA	NPKPPEDEAKEPLDPEMKLKEDPGLMPQDDSERLKVLLCQGFDDGYCLPRGFCHNGLVFKVFNCCPFSVLLKCCVR	0.971
PMBD30	MKTAFVLFALAVFLVQAMA	KPNPDVLAEDDAQMPAEDVPRPRNAMVCNAMGGNCRPGCYANEKSVGKCFASMYCCVGFQ	0.8611
PMBD31	MKFICLFFAVVFLGRAQA	DEADVQGEQDLSGPQDQNPALPAGYDEALKNEEVPLSNPIVCNSMGGRCRHKCGLSEKYVGRCFATMSSCVRFQ	0.9753
PMBD32	MKFICLFFAVVFLGIAQG	DEADLDQAEKQEALEEDLSGPQDENPDGYDEALEDEGNMQRKDRTRCQWRSRGGRCFLALCPRGTTTRIGKCTFSYLCKGEVCVPH	0.9524
PMBD33	MKFICLFFAVVFLGIVQS	DEADLDQAEKQEALEEDLSGPQDENPAGYDEALEDEGNMQRKDRTRCQWRSRGGRCFLALCPRGTTTRIGKCTFSYLCKGEVCVPH	0.9406
PMBD34	MKVQWFLAVFSGMFLVSEILVLG	GHRINRYFRSRRRKGGHCHFLKCPSTWVSGSCNIIHRCCKR	0.6199
PMBD35	MKVQWFLVVFSGLCLVSA	DITTTKDCVSFHSKCAKTCGAHAAQVGTCDGGLICCRPTG	0.9531

PMBD36	MKTVYFFYMFVALLIPNGFT	DVSDEESCRYGDVPGHCVLKKCPDGYEDIGNCGGKLRCCHYCKNKDPIAGSDFWPSL	0.9014
PMBD37	MKTVYFFYVIFALLIPNGFS	DVSDKESCVYGDVPGHCVSKECPDVYKDIGNCGGKLRCCHYGKNKDPITGSDFWPSL	0.9484
PMBD38	MKVQWLFLAVFSGGLVSA	DITTKKCEDFHSYCVKTCGAHSAQVGTGCGGLKCCRATG	0.9474
PMBD39	MKTVYFFYVIFALLIPNGFT	DVSDKESCVYGDVPGHCVSKECPDVYEDIGNCGGKLRCCHYRKNKDPITGSDFWPSL	0.902
PMBD40	MKTVYFFYMFVALLIPNGFT	DVSDEESCRYGDVPGHCVSTECPYGYEDIGNCGEKLRCCHYRKNKDPITGSDFWPSL	0.9006
PMBD41	MKTVYFFYVIFALLIPNGFT	DVSDEESCMYADLPAYCVMKECPPNLKDIGNCGEKHRCIFYRKNKYIPITGSDFWPSL	0.901
PMBD42	MKTVYFFYIAFALLIPNGFT	DVSDEESCMYGDVPGHCVTQECPDGYKDIGNCGEKLRCCHYGKNKDPITGDFWPSL	0.8985
PMBD43	MRILYFFVAVVMLLFIQCPGYA	QGPPDLDDTLACRAKQSYCIFGPCPTTFSVSGNCHGGMNCCCTK	0.9402
PMBD44	MKTCYFFLLAIALILSADPGTVFA	QVIGEELKCGEMGGACKDSCEKEYEDIGECSTTRCCIR	0.8975
PMBD45	MKTWYLPALVLLLVSVLFFDPSFA	DQASEKLCHSLKGYCENEQRCPKYVAYGICNEKMTCCIR	0.9761
PMBD46	MKTWYLLVLVFFSDPSFA	NVTDLPTCRAFRGHASYNCGDYVAIGTCAEQWVCCLR	0.9832
PMBD47	MKTWYLLVLVLFSDPSFA	NVTDQTMCHAFQGHCAVLNCHGDFVAIGICGEKWVCCLR	0.9867
PMBD48	MKTWYFVLVLFSDPSFA	MISLQILCPAFGGHCALLRCPDAFEAVDICSERWVCCVR	0.9331
PMBD49	MKTWKLVLVLLFSDPSFA	NIVDQTMCSQFSGGYCSVSECNFMTLGRCEPSEESVCCLR	0.9899
PMBD50	MKILYLLFAVFLVFIQIIPGRA	GYPADDTMECLARGFCYQCPPRTEHTGGTCQDGRLLCCKR	0.886
PMBD51	MKTLFLIFAIVLISSQA	VPGDAQAPEDIACGRGGGQVACRPGFQNVGTCKGGTMSCCRW	0.5256
PMBD52	MKICYILFCVFFLVLFIEPGFT	MINNRRKCVRHKGHCVRPTEGACKYPAFLIGRCTRKKICCKK	0.8775
PMBD53	MKFFGLLFAVLLISWASPVS	KVYGSHDCISINKGKCFRWDKCLPFYDMLGKCDLKIICCKR	0.5819
PMBD54	MKFFGLLFAVLLISWASPVS	KVYGSHDCISINKGKCFRWDKCLPFYDMLGKCDLKIICCKR	0.5819
PMBD55	MQAFWVIPLFVLRLLTSPGLS	AVRDLDEISCLYVKNLDCQKHCLYALNGLPCSEGRKCKR	0.4952
PMBD56	MKISIFYIIALSGMLLASPGSA	AWPNKAETSEDCRKAEGFCTDEPKCNTPQPDLTGTCGNGTLCCTA	0.5366
PMBD57	MQISIFYIIALSGMLLASPGVHA	GEPRCEAIGGTCGTQICERIVRSAVCPETEICVCP	0.9328
PMBD58	MKISIFYIIALSGMLLASPGHA	VRTEPIQSTEDCSAIKGFCTNEPECNTPYPVQGTGEGTLCCLP	0.8924
PMBD59	MKISIFYIIALSGMLLASPGSA	AWPNKAETSEDCRKAEGFCTDEPKCNTPQPDLTGTCGNGTLCCTA	0.5366
PMBD60	MQISIFYIIALSGMLLASPGVHA	GDTRTCFEVGGKCKSKCEKVVSDATCLTGVPCCAQ	0.9461
PMBD61	MQISIFYIIALSGMLLASPGHA	VRTEPIQSTEDCSAIKGFCTNEPECNTPYPVQGTGEGTLCCLP	0.8899
PMBD62	MQADILIIIVALSWMLLACPGYA	GSTIGEAIESAECVKDGGECTSAPNCFLTPDSPPVSSIGTCGEGTLCCVP	0.8635
PMBD63	MKSLHHSVLFVSVLLLVDPGFS	RGILKKTDCHESSGSCFPITCPWPFANHIGECIWPILRCCLYNSRST	0.791
PMBD64	MKSLHHSVLFVSVLLLVDPGFS	IKISKRKDCMRARGRCFLVCPWPLGYYIGECFWPIRRCMF	0.949
PMBD65	MNSFFLSVFLSILLADPGFS	TIRNEKHICYAQKQCKPRECDWPSSYYGGMCHMYTIVCCLPTK	0.8336
PMBD66	MKIFCAFSLMLFIALMVAPGFT	RYLHGDTECMVARGICRHLPCQPYGKKIGHCLLDTYCCKEYVMK	0.5895
PMBD67	MQLRWLFMVALLFSGMITFS	AGINGRFCERYGGSCISKNGNCPKEKHIIKSDCPSGQVCCT	0.492
PMBD68	MRLHLVFGFLCMMMMLVVPGFS	RIGNAQCREAHGKCVRRRCPGDYKRIGNCARKIACCR	0.5594
PMBD69	MRLRWLFVALLFSEMMLFMSAG	QYNKKEGEAKCENIGGFCDEEYCPSDHRQKMLCYEKAPCCVPLK	0.5439
PMBD70	MSLRWLFMVALLFSGMLLS	SPGKDYIIGKEACIKEGGHCAGEFCPSGHTLKLCDYDNPCCIPS	0.4003
PMBD71	MHLHWLFMVTLFSEILLFVSG	ETENEKLRGIQACEAKGGRCADTRCPEGSVEIKVCYLLSPCCEPTK	0.3125
PMBD72	MKLHWLFMVALLFSGMIVFSEG	SRGKACKQLGGSCIPTMYKKNCRTEKEYIEGSDCEENQVCCMKGKTKCEQGGNCISQTKCPSNKYISKTDGSSQMCCVKK	0.5519
PMBD73	MKTFAPLVIFMLLSFPGPTPTLN	ESCRGLCKPSCSGREYLASSTFCHHPDHVYCCCKTGNVSIFS	0.4682
PMBD74	MKTSAPLLLLLVLLSVPGSTP	TLNDNCSGHCSQSCSYREYVASLDRCELPGYVFCCKRIGITIFFAL	0.3919



PMBD75	MKIFAFLTAVLFFVLMAAPALA	WPKTYSECYRAHGSCHHSCPHHTRQIGECARHVRCK	0.9131
PMBD76	MKIFALFTAVLFFVLMAAPALA	QKSEAECHRLGGSCHNGFCPIGKIHKGHCGNPKRWCCRK	0.8969
PMBD77	MKIFAFLAAVLFFVLMAAPALA	QKSEAECHRLGGSCHNGFCPIGKIHKGHCGNPKRWCCRK	0.8952
PMBD78	MKFSYLISVFFLFSLLRDSADA	STPSHEPQSDEECDKAKGKCMLEICYTGWTKIGTCPHRKCCRPL	0.9559
PMBD79	MKLLHLIFSALLTLALASPGFA	NVPDENLRCVLLGGHCNYSYCRPMPRQIGICKNGSRCCKW	0.8225
PMBD80	MKLLHFLLSALLIMVLPSPGSA	TAPSSDLECQQKGLCFPGRCRPPWRSIGTCNVVLHHCKR	0.6507

*Signal peptide cleavage sites on the Beta-defensin showing signal sequence and mature peptide sequence. The probability shows the likelihood that the cleavage site is in that position*

A1.4 *Lacerta agilis* exon positions

GENE	EXON 1		Length (bp)	EXON 2		Length (bp)	EXON 3		Length (bp)	Orientation	Total Length (bp)	No of AA
	FROM	END		FROM	END		FROM	END				
LABD1	62029	62086	58	60986	61107	122				-	180	60
LABD2	66873	66930	58	68912	69042	131				+	189	63
LABD3	75487	75568	82	74104	74306	203				-	285	95
LABD4	80287	80344	58	78639	78781	143				-	201	67
LABD5	94433	94490	58	88841	88974	134				-	192	64
LABD6	102861	102912	52	99870	100024	155				-	207	69
LABD7	119806	119866	61	117048	117223	176				-	237	79
LABD8	130686	130743	58	131441	131589	149				+	207	69
LABD9	132915	132972	58	135788	135906	119				+	177	59
LABD10	147054	147111	58	143348	143460	113				-	171	57
LABD11	167232	167289	58	166137	166250	114	145451	145453	2	-	174	58
LABD12	179309	179366	58	178218	178336	119	164874	164880	6	-	183	61
LABD13	187601	187667	67	192256	192383	128				+	195	65
LABD14	242721	242778	58	240020	240156	137				-	195	65
LABD15	260964	261114	151	262307	262464	158				+	309	103
LABD16	273400	273535	136	272325	272485	161				-	297	99
LABD17	280842	280941	100	283231	283400	170				+	270	90
LABD18	290855	290984	130	292972	293141	170				+	300	100
LABD19	302639	302744	106	306548	306678	131				+	237	79
LABD20	309477	309612	136	311243	311373	131				+	267	89
LABD21	314857	315031	175	316854	316999	146				+	321	107
LABD22	334817	334982	166	327847	327992	146				-	312	104
LABD23	339625	339790	166	341725	341870	146				+	312	104
LABD24	358412	358469	58	361056	361189	134				+	192	64
LABD25	365926	365983	58	367943	368074	132	368540	368553	14	+	204	68
LABD26	380288	380345	58	381143	381294	152				+	210	70
LABD27	389131	389188	58	390242	390381	140				+	198	66
LABD28	411300	411357	58	413871	414013	143				+	201	67
LABD29	427051	427108	58	430758	430885	128				+	186	62
LABD30	450134	450191	58	448970	449088	119				+	177	59
LABD31	458835	458898	64	458296	458423	128				-	192	64
LABD32	471665	471728	64	469038	469159	122				-	186	62

LABD33	477148	477193	46	476067	476191	125				-	171	57
LABD34	482045	482090	46	480948	481078	131				-	177	59
LABD35	488025	488070	46	486911	487041	131				-	177	59
LABD36	498000	498045	46	496108	496232	125				-	171	57
LABD37	513680	513737	58	514508	514632	125				+	183	61
LABD38	547687	547744	58	549656	549792	137				+	195	65
LABD39	556577	556634	58	559600	559730	131				+	189	63
LABD40	639196	639253	58	640686	640816	131				+	189	63
LABD41	798929	798986	58	796453	796583	131				-	189	63
LABD42	806208	806265	58	804353	804480	128				-	186	62
LABD43	829061	829115	55	827030	827160	131				-	186	62
LABD44	836116	836182	67	835338	835480	143				-	210	70
LABD45	847032	847098	67	846237	846382	146				-	213	71
LABD46	855750	855807	58	853924	854048	125				-	183	61
LABD47	864716	864773	58	866310	866434	125				+	183	61
LABD48	893650	893701	52	894140	894303	164				+	216	72
LABD49	913831	913882	52	914297	914433	137				+	189	63
LABD50	933522	933573	52	933988	934124	137				+	189	63
LABD51	1040172	1040235	64	1042394	1042533	140				+	204	68
LABD52	1043546	1043603	58	1044882	1045000	119				+	177	59
LABD53	1049258	1049315	58	1050196	1050314	119				+	177	59
LABD54	1055133	1055193	61	1053883	1054028	146				-	207	69
LABD55	1062438	1062495	58	1058665	1058834	170				-	228	76
LABD56	1119000	1119063	64	1121337	1121473	137				+	201	67
LABD57	1124126	1124189	64	1125114	1125244	131				+	195	65
LABD58	1132000	1132063	64	1132986	1133119	134				+	198	66
LABD59	1140404	1140461	58	1142587	1142708	122				+	180	60
LABD60	1182392	1182449	58	1185692	1185822	131				+	189	63
LABD61	1219154	1219211	58	1219686	1219819	134				+	192	64
LABD62	1228277	1228334	58	1226968	1227107	140				-	198	66
LABD63	1370903	1370960	58	1370179	1370318	140				-	198	66
LABD64	1396051	1396108	58	1392438	1392592	155				-	213	71

*Positions coordinates of exons. Last codon of CTSB is a marker for the start of the beta-defensin cluster. Orientation of the gene and length of exons and full length of coding sequence.*

A1.5 *Lacerta agilis* Physical properties

	p/	Net Charge	Mr		p/	Net Charge	Mr
LABD1	8.69	3	4789	LABD33	6.01	-1	4468
LABD2	7.76	1	4611	LABD34	7.78	1	4546
LABD3	4.95	-5	7761	LABD35	5.48	-1	4596
LABD4	5.44	-3	5305	LABD36	6.03	-1	4329
LABD5	9.69	8	5112	LABD37	5.32	-1	4249
LABD6	7.8	1	5268	LABD38	8.33	2	4820
LABD7	8.33	2	6535	LABD39	8.36	2	4148
LABD8	8.94	4	5391	LABD40	6.91	0	4525
LABD9	7.76	1	4198	LABD41	8.98	4	4791
LABD10	5.4	-1	4014	LABD42	9.12	5	4644
LABD11	8.59	3	4285	LABD43	5.69	-1	4667
LABD12	4.68	-1	4160	LABD44	9.22	6	5219
LABD13	8.64	7	5071	LABD45	7.57	1	5254
LABD14	5.47	-1	4792	LABD46	9.49	6	4511
LABD15	5.08	-4	9697	LABD47	9.18	5	4491
LABD16	8.62	3	9152	LABD48	5.11	-1	5632
LABD17	5.26	-2	7932	LABD49	5.58	-1	4510
LABD18	4.82	-5	9130	LABD50	5.58	-1	4490
LABD19	5.63	-1	6506	LABD51	8.83	4	5381
LABD20	5.73	-1	8142	LABD52	8.63	3	4191
LABD21	4.1	-11	9242	LABD53	8.59	3	3865
LABD22	4.57	-7	9547	LABD54	7.8	1	5195
LABD23	4.68	-1	9455	LABD55	9.01	5	5899
LABD24	7.77	1	4748	LABD56	8.3	2	4887
LABD25	3.96	-8	5251	LABD57	5.09	-2	4947
LABD26	4.93	-1	5256	LABD58	5.64	-1	5084
LABD27	4.63	-4	4731	LABD59	10.09	11	4225
LABD28	7.78	1	4592	LABD60	10.08	11	4851
LABD29	4.87	-1	4066	LABD61	6.88	0	5279
LABD30	4.33	-4	4045	LABD62	8.64	3	5403
LABD31	8.62	3	4723	LABD63	8.95	6	4916
LABD32	7.52	1	4555	LABD64	9.02	5	5456

### A1.6 *Lacerta agilis* signal peptide prediction

	Signal Peptide	PRO - Mature Peptide	Probability
LABD1	MRMFLLLCVLLLLFLCQSAA	APGDLYDSLQCHYNHGHCRRLCFHNERPIGTCTNGRQRCCR	0.7311
LABD2	MNIFCLLFAGLFLVFLPNSGIT	DFVTLGCFVRGGKCETDICKENEEQIGNCSRTEKLCCKPK	0.4506
LABD3	MQASVFLLLLLLLLLLHHLISITQA	APGIIQDEKPGCDSLHHNCRVGYCSEDEIPSGGFCFEPVIICRSLPKKYKSSEETQEVAPFGDVLRLNF	0.955
LABD4	MKSFHLLVALILAVLLVSPGNG	EREPRYVSHCLRRGGICRYDDCSEGEEQIGTCYHHTMICCRDEVV	0.9438
LABD5	MRFRNLLIIAILGTFLVSLGAG	LNLFARRCRRRAGKCRGNRCFYNEIESTCYHTKIKCCREKD	0.4589
LABD6	MPSLFPVAFLLLCTLTPGHSHA	RDTLKCHKDRGTCHPTLCPAQKIEKGSCYDGIQLCCVGELVHRIVEL	0.7429
LABD7	MMFPYLSLAVQVVLVFNITAA	HRQGEEVKNAQVIICNEGRGYCLDVQFRCPSGLMFNNNNCPNKTMNKCCTPFEGRGV	0.5948
LABD8	MRLLHLSFAVVVLAFSVAA	GRAHMGSLRCPNGGRGYCLEVKFQCPSGLQFINNTCPNPKMFSCCTPVL	0.3138
LABD9	MRFLYLSFALVFLFHVAVAG	QPPSCEEQGGYCVPLTLKCPYGYIPANCGFNGNCKSK	0.7986
LABD10	MRFLYLSLALVFLFHVVEG	HFFCELYGGYCQFPATTNCTYGSRELPCGSNGICCNG	0.9919
LABD11	MKFLYLSLALVFLFHVVEG	QVRCRMFGGYCRFPETTNTYGCRRFCGSNGICNDP	0.9831
LABD12	MRFLYLSLALVFLFHVVEG	QLDCESLGGTCQLPATTNCTYAEIRWLSCGSNGICCYGK	0.9934
LABD13	MRFLSLFLLPLLLFALASQA	EKVGKVCERMRFVHSAHCPNESLPEFCGEKRKCKKLDSDA	0.8792
LABD14	MQILSLLFTLLVLLAQVATA	QSYTCFQHGVCVPSGDDCLDSGEVVPVDCGINLSCCKGKPKWR	0.6745
LABD15	MKFVCLFFALVFLCSA	QDDEAGLAKMEKQEGEDLKDLSSLQEEEDPAGDDQDSGPKASTRLAVVGCNGSRGYCLPRGYRCHNGLKWEKWNNPCFRHALLCCVR	0.94
LABD16	MKTLHFIFIVFLILHSQA	RDLWSTGELKDSKDPETERAAQADENGQTVAPPRDGKISCLWPWGYCLLRKLCASGFVMKERFNNCPNTRTLKCCVL	0.8396
LABD17	MKMLYVLFVAVFLVFAQQA	QLKPPSEDEAKLKEDPGLKQPDDSERLKVVLCNGVDGYCLPRGYSCHTSLVFEAYNDPCPSAVLKCCER	0.9359
LABD18	MKILYVLFVAVFLVFAQQA	SPKPPSKVEGEDKELLESHTYKDDGPGLNPEPKDSSRFQVVLCSNDNGYCLPRDFQCHNGLAFKEPWNDPCFSDVLKCCVR	0.9505
LABD19	MKTAFVLLALAFVFAQATA	KPNPDVLAEDDAPMPAEDVPRPRNSMVCNVMGGICRSGBKPEHKTGKCFADMVCCLPFQ	0.8639
LABD20	MKFACLFFAVVFLGIAQS	DEADVRRKQQRMPLEDLPDLRVQDPALPEDVPLPKHPIVCNTMGGRCRNWCGENERFVGRFCFIVSCCVRFQ	0.8514
LABD21	MKFTCLFFAMVFLGSAQA	DEADVQDQEEQEGDPSEDLSGPLNQNPASPAGYDGALKDEGNMQWSNGTGQCKSGGHKYFFGICPCGSVPPGKCSSFESCCTGEVCVPH	0.9715
LABD22	MKFVCLFLAMVFLGIAQS	DEADLDQAEKQQAEEALDLSGPQDENPAGYDEALEDEGNMQRKDRTRQCRSDGGKCFFAFCPRGTTRIGKCTLSYLCKGEVCVPH	0.8444
LABD23	MKFVCLFLAMVFLGIAQS	DEADLDQAEKQQAEEALDLSGPQDENPAGYDEALEDEGNMQRMNQTSQCESAGGKCFFAFCPRGTTRIGKCGRFHRCCKGEICVPH	0.8504
LABD24	MKTIYFFYIMFALLIPNPEKG	RDPVRNQRQCEKTGAFQNTPCEMGQTYTGKCADNLNCCYTV	0.5578
LABD25	MKTIYFFSVMLALLLIPYPGFT	DVSDEESCRYGDVPGYCVLKECPYGFEEIGSCSEELRCCYELQIA	0.8593
LABD26	MKTMYFFSVVLLALLLIPNPGFT	EVSQDQKSCMYGAVPAYCVMKECLIYSIDIGSCGNDIRCCYSGKNKDPI	0.8788
LABD27	MKTIYFFYIVFALLLIPKPGFT	DVSDEESCLYGDAAGHCVLKECPYHYADIGSCGGNIRCCYNGKN	0.918
LABD28	MKILYFFVAVVVLVFLQICPSSA	QGPPDLGDTIACRAKQSYCIYGSCPPTFSVSGNCHGGTCTCK	0.9176
LABD29	MKACYFLIAVLAALVSIISNQGIYA	TVNNAAEIQAGGVCTGACSSRPFNILGNCDEEKMCCKQ	0.8971
LABD30	MKTCCYFLAIALTLADPVFA	QVIGEEKCNLSGVCKDSCEADYEDIGECSTTRCCIR	0.928
LABD31	MKTWYLPALVILLVLLVFFDPSFA	DPESEKLCHSLKGYCEKELRCRPKYVAYGTCNEKRICIR	0.9834
LABD32	MAMKTWYLPVLLVLLVLSIASS	SVDTSEELCYDLGGQCRYLKCRRSPLVLLGTGCGWVWCCIR	0.4507
LABD33	MKTWYLLVLLVLFSSHPSFS	NETQQTLCHVFGGHCAVLKQCQGDYVAIDYCSERWVCCMR	0.9099
LABD34	MKTWYLLVLLVLFSSHPSFA	DIVGQIMCHSFGGYCSVSKCYADFIAVGKCVPTERWVCCLR	0.9836

LABD35	MKTWYLLVLFVSHPSFA	DVIGQTMCHNFGGYCSVSECYTDFAVGGKCAPTERWICCMR	0.9778
LABD36	MKTWYLLVLFVSHPSFA	NFALETLCHAFGGHCALLKCPSAFEAVDNCNERWVCCVR	0.8863
LABD37	MMKTCHELLLSLIIFVPGKT	DVTNTQECQLFSGICVRYICPSPHFNIGVCGPNMVCCVT	0.5623
LABD38	MKILYLLFAVFLVFIQIIPGRA	GYPADDTMECRLSRGFCKYGHCPPRTEHTGGTCQDGRLLCCKR	0.8852
LABD39	MKTLFLIFAIVLISYQAVPGDA	QAPEDTIACGRGGGGCQVGCACRPGFQNVGTCRGGTMSCCRW	0.6227
LABD40	MKVFWVIPLFLRLTLTSPGLS	ALRDLDEFSCLYVKNTDCLKHCPVHALSLGPCSAGKECCQR	0.4533
LABD41	MKLLRFLLSALLIMVLHSPGSA	TPPSSDLECCQQKGLCFPGRCRRPWRSIGTCNVVWHRCCER	0.9607
LABD42	MKLLHLIFSALLIALASPGFA	NVPDENLRCILNGGYCNYCYRRPMPRQIGICKNGTRCCKM	0.7851
LABD43	MNIFFIYALLVFSLLEDPA	EKITTEICHENGKCAAYMECRDNAKEIGKIDPLYLCCKD	0.834
LABD44	MKFSHLFSVFFLSSFLLRDSAHA	FKPTYPPQSKKECKAEGKCMRERCMTAWKRIGKCPNDVCCCLP	0.9899
LABD45	MKFSHLISVFFLSSFLLRDSAYA	SKPSHEPESDEECEKAKGKCMMEFCYSSSWKIGTCKPHRVCCQV	0.9803
LABD46	MKIFALAAVFFVFLMAAPALA	QTSERECRRMGGSCHNGFCPPGKHHRGHCGNQKIWCCRR	0.8843
LABD47	MKIFALFTAVLFFVFLMAAPALA	QTSERECRRMGGSCHNGFCPHGTYHKGHCNPKIWCRR	0.8904
LABD48	MKTSASLVLLLVLVSPGPATN	NNILCNVCKRRCSAGEYKSSLDPCQESGYVFCCKEENVLNCCYELQSL	0.3743
LABD49	MKTSAFVLVLLLLLVPGPNT	QDLCDGVCRRRCTAGEYESPLDPCSESGHVFCCKKFGNVLN	0.9364
LABD50	MKTSASLVLLLLLVPGPNT	QDLCDGVCRRRCTAGEYESPLDPCSESGHLICCKKFGNVLN	0.9423
LABD51	MQLRWLFMVALLFSGMIMISA	GDKERKGETCKRNGGFCVRSKETCPSKDYINLYNDCPTGQQCCKKQ	0.6543
LABD52	MPLRWLLMLLSVLLPFSEG	NSKGAGCKTIGGVCMPKNNCQKLYTESDCGKHEVCCQK	0.768
LABD53	MPLRWLLMLAISMLLFFSGETAKG	TGCKTIGGECKPKKCKNIYTESDCPKDQVCCCEVK	0.7043
LABD54	MMPLSWLFLMLSLMSAMFFSTGDA	AIKPAHEPAVRACETDGGYCDGEICPRGYSRLRMICYDKIPCCIRY	0.9436
LABD55	MPLCWLFMLAVFVLLLFSTGDISA	YDPAVENAIQKGKEECKRKRGFCEGHFCSTGTKKDGECYSHISCCVRKKNKT	0.9128
LABD56	MSLRWLFMVALLYSEMILLSSA	GKNYMIKKEACITEGGHCAGEFCPAGHREIKLCYGNVPCCVPSKS	0.7196
LABD57	MRLRWLFVLALLFSGMILLSSA	AQYNKKEGENACENIGGFCDGEGYCPIDHRQKMFCEYEGAPCCVVK	0.6467
LABD58	MRLRWLFVLALLFSGMILLSSA	GQDNRKEIETLCKNIGGYCNEEYCPNQQIKKMSCYEGAHCCVPLN	0.5993
LABD59	MRLRHIVFALLCLMMLVVPGLS	KIGNAQCRDAHGKCVKRRCTGGYKRIGNCNKKVACCR	0.6978
LABD60	MRNSYLALVLLFLGLFLAPGLC	TGIRKAKDCKKAHGRCCRAMCLHDPWKRIGKCDFKRFCCVK	0.5536
LABD61	MKILPILSVVFLSFLVAP	GFAHIGIYSREQCEYFKGRCVLFQCEEHWRKVKGCAADVCCSQE	0.9041
LABD62	MKIFCALSLMLFIELMVAP	GFTRYLQGDTECMVARGICRHLPCQPHAKKIGHCLLNTYCCKEYVMN	0.3906
LABD63	MQFGIVYIVAVSWMLLAFPGYT	CVKDQPKNEKACTDKGWVCSAKDKCPATHLKKIKCAIDRFCCAM	0.6653
LABD64	MQFGVIYIVALSWMLLTFPGYA	YGRPKPKNEKPKNENSCKAKGGDCTTNEECQGLKSHGIKCDKNRICCIM	0.5148

*Signal peptide cleavage sites on the Beta-defensin showing signal sequence and mature peptide sequence. The probability shows the likelihood that the cleavage site is in that position*

A1.7 *Zootoca vivipara* exon positions

GENE	EXON 1		Length (bp)	EXON 2		Length (bp)	EXON 3		Length (bp)	Orientation	Total Length (bp)	No of AA
	FROM	END		FROM	END		FROM	END				
ZVBD1	37719	37779	61	36678	36799	122				-	183	61
ZVBD2	41173	41230	58	43231	43388	158				+	216	72
ZVBD3	51869	51938	70	50409	50611	203				-	273	91
ZVBD4	57392	57449	58	56136	56278	143				-	201	67
ZVBD5	65688	65745	58	62745	62878	134				-	192	64
ZVBD6	71680	71731	52	70488	70642	155				-	207	69
ZVBD7	93351	93411	61	91303	91478	176				-	237	79
ZVBD8	101751	101808	58	103534	103649	116				+	174	58
ZVBD9	172604	172661	58	174275	174393	119				+	177	59
ZVBD10	191735	191801	67	195686	195813	128				+	195	65
ZVBD11	249793	249838	46	250790	250914	125				+	171	57
ZVBD12	287246	287309	64	287716	287843	128				+	192	64
ZVBD13	296757	296814	58	295714	295814	101				-	159	53
ZVBD14	302681	302732	52	305039	305178	140				+	192	64
ZVBD15	310257	310314	58	311316	311416	101				+	159	53
ZVBD16	343675	343732	58	343102	343244	143				-	201	67
ZVBD17	369922	369979	58	368862	368993	132	368345	368355	11	-	201	67
ZVBD18	377696	377753	58	376600	376731	132	376104	376114	11	-	201	67
ZVBD19	387050	387107	58	386072	386203	132	385124	385131	8	-	198	66
ZVBD20	459023	459188	166	457211	457356	146				-	312	104
ZVBD21	522077	522182	106	517265	517395	131				-	237	79
ZVBD22	534768	534897	130	533415	533584	170				-	300	100
ZVBD23	540692	540815	124	538781	538950	170				-	294	98
ZVBD24	545016	545151	136	546054	546214	161				+	297	99
ZVBD25	559719	559860	142	557471	557628	158				-	300	100

ZVBD26	572323	572437	115	573846	574003	158				+	273	91
ZVBD27	697481	697538	58	698395	698525	131				+	189	63
ZVBD28	704367	704424	58	703829	703962	134				-	192	64
ZVBD29	754895	754952	58	755427	755560	134				+	192	64
ZVBD30	775705	775768	64	774202	774335	134				-	198	66
ZVBD31	786820	786883	64	785077	785210	134				-	198	66
ZVBD32	826759	826816	58	829885	830069	185				+	243	81
ZVBD33	908130	908187	58	909477	909604	128				+	186	62
ZVBD34	915066	915123	58	918939	919069	131				+	189	63

*Positions coordinates of exons. Last codon of CTSB is a marker for the start of the beta-defensin cluster. Orientation of the gene and length of exons and full length of coding sequence.*



A1.8 *Zootoca vivipara* Beta-defensin physical properties

	pI	Net Charge	Mr		pI	Net Charge	Mr
ZVBD1	8.67	3	4787	ZVBD18	5.57	-1	5036
ZVBD2	7.75	1	5491	ZVBD19	4.36	-2	5040
ZVBD3	6.92	0	7719	ZVBD20	4.41	-9	9348
ZVBD4	5.44	-3	5305	ZVBD21	6.97	0	6401
ZVBD5	9.3	6	5172	ZVBD22	5.77	-1	9115
ZVBD6	6.98	0	5277	ZVBD23	5.26	-2	8629
ZVBD7	8.32	2	6466	ZVBD24	5.31	-1	9025
ZVBD8	8.24	2	4163	ZVBD25	5.33	-2	9318
ZVBD9	6.03	-1	4273	ZVBD26	7.76	1	8358
ZVBD10	9.22	6	5041	ZVBD27	8.65	3	4689
ZVBD11	3.92	-3	4171	ZVBD28	6.88	0	4952
ZVBD12	8.31	2	4661	ZVBD29	8.83	4	4854
ZVBD13	4.29	-4	3865	ZVBD30	5.66	-1	5079
ZVBD14	8.92	4	4973	ZVBD31	7.89	1	4833
ZVBD15	4.29	-4	3865	ZVBD32	9.33	7	6494
ZVBD16	7.78	1	4599	ZVBD33	9.15	5	4586
ZVBD17	4.96	-3	5116	ZVBD34	8.98	4	4763

*Physical properties of the Zootoca vivipara mature peptide beta-defensins. All properties were achieved by using the protparam program on the ExPASy Server (Gasteiger, E. et al. 2005)*

### A1.9 *Zootoca vivipara* signal peptide prediction

	Signal Peptide	PRO - Mature Peptide	Probability
ZVBD1	MKMLYLLYGVLFLFLYQSAA	APGDLYDSLQCHYNHGHCRRLCFYNERPIGTCTNGRQRCCCK	0.7266
ZVBD2	MKIFCLLFAGLFLVFLPNSVML	LFPPVLVGIADFVTLGCFVRGGKCEDICKENEEQIGNCSKTEKLCKKPK	0.2814
ZVBD3	MQTSVFLLLLLLHHHTSITLA	APGIIQDGKPGCDLHHCRCRMGYCSENEIPSGGFCFKPVIICCRSLPKKYKSSSEETQEVAPFGDVLRLNF	0.9245
ZVBD4	MRSFHLLVALFLAVLLVSPGNG	EREPRYVSHCLRRGGICRYDDCSEGEEQIGTCYHHTMICCRDEVV	0.9389
ZVBD5	MRFRNLLIVAILGAFVLSLGGAG	LNLFVRRCWRRAGKCRGNRCFYDEIEISTCYHTKIKCCREKD	0.576
ZVBD6	MPSLYLIVLFLCTLTPGSHA	RDTLKCHEDRGTCHPTLCPAQKIEKGSCYSGFQLCCVGLVHRITEL	0.717
ZVBD7	MMFPYLSLAVQVVVFNIAA	HRQGDEVKNLAQVLCNEGRGYCLDVQFSCPSGLMFNNKCPNKT MNKCCTPFEGRGV	0.6715
ZVBD8	MRFLYLSFALVFLFHAVAG	QQSCKEQGGYCVPLTEKCPYGYIPANCGFNGNCCCK	0.6964
ZVBD9	MRFLFLSALVFILSHVVA	ENPLCQFFGGTCHFPATTKCAHGEWTGNLCGPNGVCCRSE	0.9561
ZVBD10	MRFLSLFLLPLFLALASQA	EKGSKDCKRMRGFCVHKSACPSNTILPFKCGDKQKCKKLDADV	0.8422
ZVBD11	MKTWYLLVLVLFSDPSFA	ALAMQTLCYSGGQCTYLNCPAAFEAVDDCSESSVCCIR	0.723
ZVBD12	MKTWYLPALFLLVSVLFFDPSFA	DPASEKLCHILNGYCENEVRCRPKYVAYGICNEKRICIR	0.981
ZVBD13	MKTCYHFLLAIALILSA	DPGEEKCNFSGGVCKDSCESGYDDIGECSTTRCCIR	0.3583
ZVBD14	MKTCYLLFALGFLFHPGLP	ANVYNRSQCREWNGVCAFYKCPATFNSIGKCLTFRPCCLLQLPG	0.7265
ZVBD15	MKTCYHFLLAIALILSA	DPGEEKCNFSGGVCKDSCESGYDDIGECSTTRCCIR	0.3583
ZVBD16	MRILYFFVAVVLLFQICPGYA	QGGPPDLGDTLACRAGQSYCFGPCPPTFSVSGNCHGGLNCCTK	0.9436
ZVBD17	MKSICIFNVVLLALLIPNGFT	DVSDKESCMYGDVPGYCVLEELYGYTKIGSCGKDIHCCHYVRLY	0.9112
ZVBD18	MKTIYIFNVVLLALLIPNGFT	DVSDKESCMYGDVPGYCVLEELYGYTKIGSCGKDIHCCHYVRLY	0.9046
ZVBD19	MKTIHFNVVLLALLFIPNP	GFTGVSDLKSCMYGDVPAYCVMKECPLIYSDIGSCGNDIRCCYSGAF	0.7236
ZVBD20	MKFVCLFFAVVFLGIAQS	DEADLQGAEKQEALEDLGSPQDENPAGYDEALEDEGNMQWRDGTSKCESAGGKFFAFPCPRGTTTRIGKCSLFRCLCKGEVCPH	0.8545
ZVBD21	MKTALVLLALAFVFAQATA	KPNPGVLAEDDAPMPAEDVPRPRNAMVCNAMGGNCRAGCHSHEKSLGKCFANMHCCVAYQ	0.825
ZVBD22	MKMLYVLFVAVFLVQVQA	NPKPPSKVEGEAKELLESRAYKDDGPKPEPKDSSRFQVVLCSNNGYCLPRDFQCHNGLAFKEPWNDPCFSDVLKCCVR	0.958
ZVBD23	MKMLYVLFVAVFLVQVQA	NPKPSEDEAKEPLDPETKLKEDPGLLPQDDARLKVVLCSGIDGYCLPRGFACHNGLVFKEAFNNCPFSAVLKCCVR	0.9569
ZVBD24	MKTLHFIFVFLVFLVLSQA	GLDLWSTGKLKDSKDPETERAAYQADENGLAVAPLDEKISCLWPWGYCLLRELQCSSGFVMKERFNCPNTRTLKCCVL	0.8144
ZVBD25	MKFVCLFFALVLLCNA	QEDEADLTMEKEQGEDLKAAGDTQGLQDEDSGMKASPRLAVVLCNGNRYCLPRGYRCSNGLVFKEPWNNCPSRHALKCCVR	0.9775
ZVBD26	MKFVCLFFALVLLCNA	QEDEADLAKMEKEQGEDLKAADSGMKASPRLAVVLCNGNRYCLPRGYRCSNGLVFKEPWNNCPSRHALKCCVR	0.9742
ZVBD27	MKIFCALSLMLFIALMVAPGFT	RYLQGDTECMVARGICKHLPCQPHARKIGHCLLNTYCKDI	0.546
ZVBD28	MKILSILSVVFLVFLVAPGFT	HIGINSREQCDYFKGKCLLFCQEEHWRKVGKCAPDVYCCSQE	0.5978
ZVBD29	MKAFSSLFLLSLLLFISSPASS	EKIEKPWQCSKQKICMTLKECLLPYKPIGKCDADTHCCQKK	0.9588
ZVBD30	MRLRWLFVVALLFSEMLLFSAG	QDEKKEIETLCKNIGGYCDDKFCPSDHKEKMTVIPPRTTNGYF	0.5778
ZVBD31	MRLRWLFVVALLFSEMFLFSAGG	FNKQAGERACEKMGGFCDGEGCPSDHRQTMICYSGRGTTHRFL	0.3921
ZVBD32	MPLCWLFLMAICVLLLSAGDISA	YDPAVLNAIENGKKECRKKGFCDGHTCPVGTQNGECYHHLSCCVKRRKNNRNRKS	0.8559
ZVBD33	MKLLHLIFSALLTIALASPGFA	NVPDENLRCILNNGGHCNYSYCRPPMRQIGICKNGSRCCKL	0.7818
ZVBD34	MKLLRFLSALLIMVLPSPGSA	TPSSDLECCQQGLCFPGRCRRPWRSIGTCNAVWHRCER	0.9455

Signal peptide cleavage sites on the Beta-defensin showing signal sequence and mature peptide sequence. The probability shows the likelihood that the cleavage site is in that position

A2.1 *Crotalus viridis viridis* exon positions

GENE	Exon 1		Length (bp)	Exon 2		Length (bp)	Orientation	Total Length (bp)	No of AA
	FROM	END		FROM	END				
CVBD1	68639	68696	58	67178	67299	122	-	180	60
CVBD2	74090	74147	58	73511	73689	179	-	237	79
CVBD3	80255	80312	58	79220	79362	143	-	201	67
CVBD4	85295	85352	58	86999	87186	188	+	246	82
CVBD5	88569	88623	55	89680	89831	152	+	207	69
CVBD6	117037	117094	58	118499	118644	146	+	204	68
CVBD7	141093	141150	58	139622	139749	128	-	186	62
CVBD8	160041	160094	54	159427	159552	126	-	180	60
CVBD9	169283	169337	55	168100	168221	122	-	177	59
CVBD10	187078	187135	58	188737	188870	134	+	192	64
CVBD11	200474	200531	58	205537	205661	125	+	183	61
CVBD12	247452	247509	58	241999	242138	140	-	198	66
CVBD13	271428	271482	55	268336	268466	131	-	186	62
CVBD14	299523	299580	58	298373	298497	125	-	183	61
CVBD15	314977	315034	58	313550	313806	257	-	315	105

*Position coordinates of exons. Last codon of CTSB is a marker for the start of the beta-defensin cluster. Orientation of the gene and length of exons and full length of coding sequence. Number of amino acids in full peptide shown.*

A2.2 *Crotalus viridis viridis* mature peptide physical properties.

GENE	pI	Net Charge	Mr
CVBD1	8.3	3	4501
CVBD2	6.77	0	6225
CVBD3	5.81	-2	5322
CVBD4	8.54	4	7008
CVBD5	8.26	3	5504
CVBD6	7.62	1	4639
CVBD7	3.67	-4	4774
CVBD8	8.57	4	4312
CVBD9	9.09	6	4319
CVBD10	7.99	2	4649
CVBD11	3.93	-1	4060
CVBD12	8.56	4	4671
CVBD13	9.35	6	4720
CVBD14	8.28	3	4871
CVBD15	8.96	5	9818

Physical properties of the *Crotalus viridis viridis* mature peptide beta-defensins. All properties were achieved by using the protparam program on the ExPASy Server (Gasteiger, E. *et al.* 2005)

A2.3 *Crotalus viridis viridis* signal peptide prediction.

GENE	SIGNAL PEPTIDE	MATURE PEPTIDE	Probability
CVBD1	MKIIFMLWALFLFLCQPIPAKG	DLYDSLVCRRNNHGHCRRLCFHHEQVIGTCMNGRQRCK	0.8194
CVBD2	MKMISIIIFASIVLSFLASSGKG	DFVTLGCLFRGGTCETDTCKENEIQIGNCSKTEKICCKKPKPALHPPGEIITRSTDT	0.4895
CVBD3	MKICHLAIALFFAMLLVSPNG	ERMVRFVSHCLRRGGICRYDDCSEGEEQIGTCYHHTMICCRDEVM	0.9661
CVBD4	MSSCKLLIVALFAIFMLSLGSS	LNLSARLCWQRGGRCHRNSQCYNEIEIGTCYHFRKCCRDKSQEDGSSYVLGKVFNGFTC	0.4823
CVBD5	MRFLFLIFALLLLFYVSSG	QLRNCISDGGFCQTGLQKECEFGSLPYNCGINALCCKRGPVRAIFPLVMR	0.5134
CVBD6	MNILYCFTAVIFFFFHAAQEDRFIE	AILLTCASMGGFCILQPNETCPGILLDVPCHFRRRCCSKTDA	0.2463
CVBD7	MKIYHLFFLTLFLKGVVGP	DPAGSQLACQYRFDGFCSPEDLPCPNCFVSYGSCEFDQCCVK	0.4191
CVBD8	MKAWCVLLFLTFVILSDLGEA	VRIMDSRTCLINKGVCKKSCPDLTKIGLCHVNEPCKA	0.993
CVBD9	MKAWCVLLFLTFVILSDLGEA	KDIRNAKCNKQKGVCKKTCPPYKNGVCQINIPCCVP	0.988
CVBD10	MKTLFLLFAALLFFSQIIPGNS	QPAPDTLECRSSHGFCYQCPGETLPTGGTCQWGRLVCKS	0.8347
CVBD11	MKAFLLLVAIFLFSNQAVTA	TGQSDPANIACFQNGGSCRLSCFPFGVASGDCAGGLVCCIW	0.5898
CVBD12	MQFGILSFVLASWILLASPGCA	VRTVPIKTKADCSAIGFCTINPECNTPYPIQGTGKGTLCCLK	0.6157
CVBD13	MKTMCLIRFFFVLMVQPGAQF	IASRKQCERVGGVCIFSFSCYWPLRIRIGRCNFFLRCCSF	0.2979
CVBD14	MKMFTEFFSLIFFLALLSVSG	IKTDIWDSECIKFGGICKHWPCRPFRRIGFCVYNTYCKE	0.8001
CVBD15	MKILNLICVLFCAFLFTPGTG	NTIKTEKECHAADGYCKMGECLYPKFKLIGFCRKVFYCKKNVQRNNRYVTVHQNMEVNRIYTYILNIAGTGNNNDEDNRRNA	0.7924

Sequence of signal peptide and mature peptide. Probability and cleavage site predicted using SignalIP – 5.0 server (Almagro Armenteros et al 2019)

A.2.4 *Naja naja* exon positions.

GENE	Exon 1		Length	Exon 2		Length	Orientation	Total Length	No of AA
	FROM	END	(bp)	FROM	END	(bp)		(bp)	
NNBD1	39048	39105	58	37694	37815	122	-	180	60
NNBD2	48385	48442	58	47189	47352	164	-	222	74
NNBD3	54401	54458	58	53034	53176	143	-	201	67
NNBD4	58228	58285	58	59934	60088	155	+	213	71
NNBD5	64699	64753	55	67375	67526	152	+	207	69
NNBD6	78088	78142	55	81410	81561	152	+	207	69
NNBD7	131611	131668	58	124024	124148	125	-	183	61
NNBD8	198284	198341	58	207249	207373	125	+	183	61
NNBD9	255899	255953	55	256436	256563	128	+	183	61
NNBD10	291629	291686	58	293705	293832	128	+	186	62
NNBD11	329916	329973	58	331665	331801	137	+	195	65
NNBD12	347727	347784	58	351528	351652	125	+	183	61
NNBD13	408403	408460	58	400354	400493	140	-	198	66
NNBD14	447258	447315	58	445201	445349	149	-	207	69
NNBD15	473357	473414	58	467839	467972	134	-	192	64
NNBD16	496559	496616	58	495377	495501	125	-	183	61
NNBD17	519608	519665	58	517588	517844	257	-	315	105
NNBD18	530207	530264	58	528997	529127	131	-	189	63
NNBD19	555949	555994	46	556190	556323	134	+	180	60

NNBD20	755118	755175	58	756599	756711	113		+	171	57
NNBD21	794343	794397	55	792165	792283	119		-	174	58
NNBD22	805901	805958	58	807692	807810	119		+	177	59
NNBD23	900882	900939	58	907486	907604	119		+	177	59
NNBD24	965504	965561	58	967102	967220	119		+	177	59
NNBD25	1022204	1022261	58	1025219	1025334	116		+	174	58
NNBD26	1031135	1031192	58	1032464	1032582	119		+	177	59
NNBD27	1078767	1078824	58	1080034	1080152	119		+	177	59

*Position coordinates of exons. Last codon of CTSB is a marker for the start of the beta-defensin cluster. Orientation of the gene and length of exons and full length of coding sequence. Number of amino acids in full peptide shown.*

## A2.5 *Naja naja* mature peptide physical properties

GENE	<i>p</i> <i>i</i>	Net charge	Mr	GENE	<i>p</i> <i>i</i>	Net charge	Mr
NNBD1	8.69	5	4453	NNBD15	9.55	7	4860
NNBD2	9.1	5	5829	NNBD16	9.08	5	4879
NNBD3	5.83	-2	5326	NNBD17	8.92	5	9785
NNBD4	8.32	2	5749	NNBD18	9.16	7	4603
NNBD5	9.3	6	5419	NNBD19	9.33	7	4943
NNBD6	9.1	5	5567	NNBD20	9.15	5	3911
NNBD7	5.49	-1	4635	NNBD21	7.79	1	4376
NNBD8	5.5	-1	4681	NNBD22	9.61	7	4450
NNBD9	9.64	8	4392	NNBD23	7.79	1	4429
NNBD10	6.86	0	4532	NNBD24	9.22	5	4505
NNBD11	7.79	1	4755	NNBD25	8.94	4	4348
NNBD12	4.32	-2	4149	NNBD26	8.96	4	4220
NNBD13	7.75	1	4904	NNBD27	8.68	3	4160
NNBD14	4.74	-5	4583				

Physical properties of the *Naja naja* mature peptide beta-defensins. All properties were achieved by using the protparam program on the ExPASy Server (Gasteiger, E. *et al.* 2005).



## A2.6 *Naja naja* signal peptide prediction

GENE	SIGNAL PEPTIDE	MATURE PEPTIDE	Probability
NNBD1	MRIIFMLWALFLFLCQPVPAKG	NLYDSLVCRNIGHGCRRLCFHHEQVIGTCTNGRQHCKK	0.6504
NNBD2	MKMICIFFFASVILSFLASSGKG	DFVTLGCLFRGGTCETNTCKENEVQIGNCSKIQKICCKKPKPAPQRIATWRN	0.45
NNBD3	MKICHLAIVLFFATLLVSSGNG	ERMVRFVSHCLRRGGICRYDDCSEGEEQIGTCYHHTMICCRDEVN	0.7568
NNBD4	MSSCKLLVVALFAVFLISLGSS	LNLSARLCWQKGGRRHRAGQCYDNEIEIGMICYHFLKCCRDKSQEETPT	0.4714
NNBD5	MKFLFLTFALLLLFYVSSA	QVKTCMSDGGFCQ TGLFKPKYGS LYPNCGINGICCKRGPVRSTSLVMR	0.5129
NNBD6	MRFLFLTFALLLLFYVSSG	QPDDCVKDGGFQIGLQKRCVYGS LYPNCGINAKCKRLPVRSMFLVVR	0.7103
NNBD7	MKAFLLLVAIFLFSHQAVT	EIDESDPLVIACNMLGSCWSYCPHTTIARGRCPPGLFCCTW	0.7136
NNBD8	MKAFLLLVAIFLFSHQPVTA	DIIDDYASVHCDRRRGYCADECIPGFFFPKCFGGQTCCKW	0.9868
NNBD9	MKAWCVLFFLTFVILSDLGEA	QRIMDSRACIKNKGACRKLSCAVSEKKIGLCHVNKPCCKA	0.9933
NNBD10	MKIYHLFFLTFLKGVVGPAS	GGQHVCQQHLNGFCYPKIVSCPDCFTFYGNCFNLQCCAR	0.2488
NNBD11	MKTLFLLFAALLFFSQIIPGSS	QPAPDTLECRNQDHGFCKRYNCPGQTVHTGGTCQWGTLLCCKS	0.7524
NNBD12	MKAFLLLVAIFLFSHQAVTA	TGESDPANIACFKSGGSCRVPFPFAEQSGDCAGGLVCCQW	0.6178
NNBD13	MQLGILFVVLASWILLASPGCA	VRTERIDTEVLCKAIKGFCTHNTTECNTPYPIQGTGKDTLCLLK	0.6034
NNBD14	MQFAILSIAFISWMFLVFPDGGDDAGA	DAEEEVSSAQCMNSHGGRCIKECAEDEKEIHKCPAGVCCKEQ	0.2197
NNBD15	MKTMWHLVRRFFVLMVQPGAP	KLTVNRKQCEKAGGLCIFSFYCIWPARIKIGRCSLFVRCCTF	0.6092
NNBD16	MKMFTFFSLIFFLTLLSVAG	IKQNVWGDRECIQVGGLCXHWPCRFKRIKIGFCVYKSYCCKE	0.6065
NNBD17	MKVLYLTCVFFCIAFLFTPGIG	NTIKTEKECHAAEGYCKMGECLYPKFKLIGFCRKFVYCKKNVQRYNRYVTVHKNMEVNRIYTYILDSAGTGNKDDNDRNV	0.7163
NNBD18	MKIFGIFSAIFLALIMAIA	CKVPKTAADC DREGGKCRFLRCPNLTAIGKCDKNGGVCCCKQ	0.5185
NNBD19	MKILLVCLILFLSLSGFTQA	KPKRCPYGGVCRSYKEYCYREEKYWGRC PWNQKKIYCCFW	0.9426
NNBD20	MKVLYLVLTFLFLAILPEPGNA	NYVCKRRDGI CVRFYCPGKNLGTWGCNGLTCCR	0.9835
NNBD21	MKILYLFAFLFLAFLSEPGNA	IYECHRQRGECFRIQCPNGYQDLGTLGCPEGWRCCRQ	0.9706
NNBD22	MKILYLLFAFLFLAFLSEPGSA	QRWCRRQRGRCCYGHCLLNHRDIGRQDCGPKSKCCVP	0.972

NNBD23	MKILYFLFAFLFLAFLSEPGNA	HHLCGSQGGRCFRYQCFFGYEDLDQVDCQWRWKCCRP	0.9822
NNBD24	MKILYFLFAFLFLAFLSEPGNA	HHLCGSQGGRCHQYRCRPRHDDLGRDCPWRWKCCRP	0.9834
NNBD25	MKILYLLFSFLFLAFLSDPGNA	QRVCRGLGGRCYRDCPKNTEDIHRKDCRHEWTCCRP	0.9928
NNBD26	MKILYLLFAFLFLVCLSQPGNA	QSQCNGLRGVCYRPHCPHGLQYLGQVDCRWGAVCCRR	0.9196
NNBD27	MKILYLLFAFLFLVCLSQPGNA	QSQCGSLRGVCYRPHYCPHGLQYLGQVDCPWGAVCCRR	0.914

*Signal peptide cleavage sites on the Beta-defensin showing signal sequence and mature peptide sequence. The probability shows the likelihood that the cleavage site is in that position.*

A2.7 *Thamnophis elegans* exon positions

GENE	Exon 1		Length	Exon 2		Length	Orientation	Total Length	No of AA
	FROM	END	(bp)	FROM	END	(bp)		(bp)	
TEBD1	35007	35064	58	33730	33851	122	-	180	60
TEBD2	47352	47409	58	45000	45163	164	-	222	74
TEBD3	53307	53364	58	51979	52121	143	-	201	67
TEBD4	57808	57865	58	59122	59276	155	+	213	71
TEBD5	66037	66091	55	66443	66594	152	+	207	69
TEBD6	77374	77431	58	79754	79908	155	+	213	71
TEBD7	82404	82458	55	82810	82961	152	+	207	69
TEBD8	86906	86960	55	90467	90618	152	+	207	69
TEBD9	122538	122595	58	121439	121557	119	-	177	59
TEBD10	132826	132883	58	131734	131852	119	-	177	59
TEBD11	143442	143499	58	145545	145690	146	+	204	68
TEBD12	166123	166180	58	164529	164647	119	-	177	59
TEBD13	189274	189328	55	188668	188795	128	-	183	61
TEBD14	200583	200637	55	199977	200104	128	-	183	61
TEBD15	212604	212658	55	212057	212172	116	-	171	57
TEBD16	252369	252426	58	254023	254159	137	+	195	65
TEBD17	307298	307355	58	311767	311891	125	+	183	61
TEBD18	326082	326133	52	328378	328505	128	+	180	60
TEBD19	370243	370294	52	371918	372045	128	+	180	60
TEBD20	410515	410572	58	408685	408803	119	-	177	59
TEBD21	429691	429742	52	432015	432223	209	+	261	87
TEBD22	459026	459083	58	457331	457449	119	-	177	59
TEBD23	486449	486500	52	488886	489007	122	+	174	58
TEBD24	534158	534215	58	535368	535486	119	+	177	59
TEBD25	600720	600777	58	602384	602520	137	+	195	65

TEBD26	661129	661186	58	665635	665759	125		+	183	61
TEBD27	679024	679075	52	681320	681447	128		+	180	60
TEBD28	712302	712359	58	710471	710589	119		-	177	59
TEBD29	728458	728509	52	730133	730260	128		+	180	60
TEBD30	770870	770921	52	773188	773315	128		+	180	60
TEBD31	799117	799174	58	797165	797283	119		-	177	59
TEBD32	821670	821721	52	824096	824217	122		+	174	58
TEBD33	850640	850691	52	855921	856042	122		+	174	58
TEBD34	907499	907556	58	905481	905599	119		-	177	59
TEBD35	934282	934339	58	936388	936521	134		+	192	64
TEBD36	959598	959655	58	962670	962794	125		+	183	61
TEBD37	970410	970461	52	972717	972835	119		+	171	57
TEBD38	1214608	1214665	58	1215889	1216001	113		+	171	57
TEBD39	1258970	1259027	58	1260980	1261092	113		+	171	57
TEBD40	1331387	1331444	58	1332294	1332400	107		+	165	55
TEBD41	1367330	1367387	58	1368707	1368825	119		+	177	59
TEBD42	1375130	1375187	58	1377119	1377231	113		+	171	57
TEBD43	1400133	1400190	58	1398866	1398981	116		-	174	58
TEBD44	1431072	1431129	58	1433245	1433360	116		+	174	58
TEBD45	1454306	1454363	58	1455488	1455600	113		+	171	57
TEBD46	1501688	1501745	58	1502795	1502916	122		+	180	60
TEBD47	1545655	1545712	58	1544494	1544612	119		-	177	59
TEBD48	1612986	1613043	58	1614236	1614351	116		+	174	58
TEBD49	1650260	1650317	58	1651594	1651712	119		+	177	59
TEBD50	1672202	1672259	58	1673505	1673623	119		+	177	59
TEBD51	1697966	1698023	58	1699237	1699355	119		+	177	59

*Positions coordinates of exons. Last codon of CTSB is a marker for the start of the beta-defensin cluster. Orientation of the gene and length of exons and full length of coding sequence.*

## A2.8 *Thamnophis elegans* Beta-defensin physical properties

GENE	pI	Net Charge	Mr	GENE	pI	Net Charge	Mr
TEBD1	8.98	4	4791	TEBD26	5.92	0	4028
TEBD2	9.08	5	5868	TEBD27	4.78	-1	4744
TEBD3	5.83	-2	5326	TEBD28	10.72	10	4640
TEBD4	8.32	2	5689	TEBD29	7.76	1	4569
TEBD5	8.32	2	5614	TEBD30	8.96	4	4607
TEBD6	8.32	2	5687	TEBD31	8.32	2	3917
TEBD7	8.89	5	5539	TEBD32	3.61	-5	4406
TEBD8	8.89	4	5542	TEBD33	4.03	-4	4338
TEBD9	9.86	8	4135	TEBD34	8.34	2	4064
TEBD10	9.18	5	3847	TEBD35	8.68	2	4761
TEBD11	6.71	0	4454	TEBD36	4.21	-1	4087
TEBD12	3.9	-4	4584	TEBD37	7.76	1	4195
TEBD13	9.42	7	4373	TEBD38	5.38	-1	3590
TEBD14	9.42	7	4373	TEBD39	7.78	1	3663
TEBD15	8.33	2	3908	TEBD40	8.65	3	3542
TEBD16	8.65	3	4970	TEBD41	8.98	4	4132
TEBD17	5.92	0	4028	TEBD42	9.26	5	3951
TEBD18	4.78	-1	4744	TEBD43	9.25	5	4412
TEBD19	6.12	0	4526	TEBD44	11.71	13	4658
TEBD20	9.42	7	4241	TEBD45	9.25	5	4066
TEBD21	8.91	5	7954	TEBD46	9.13	5	4244
TEBD22	8.65	3	3945	TEBD47	9.86	8	4162
TEBD23	3.61	-5	4372	TEBD48	9.21	5	4322
TEBD24	9.25	5	4460	TEBD49	9.68	6	4262
TEBD25	8.65	3	4970	TEBD50	10.62	9	4274
				TEBD51	10.59	10	4546

Physical properties of the *T. elegans* mature peptide beta-defensins. All properties were achieved by using the protparam program on the ExPASy Server (Gasteiger, E. *et al.* 2005).

A2.9 *T. elegans* signal peptide prediction

	SIGNAL PEPTIDE	MATURE PEPTIDE	Probability
TEBD1	MRIIFILWAFFLFLCQSIPA	REELHDTLTCRNNHGRRCRRICFHHEHVIGTCTNGRQRCK	0.4521
TEBD2	MRMICIFFASVILSFLASSGKG	DFVTLGCLFRGGTCETNTCKENEVQIGNCTKIQKICCKPKPALHKTEIRRN	0.4384
TEBD3	MKICHSIALFFAMLLVSSGNG	ERMVRFVSHCLRRGGICRYDDCSEGEQIGTCYHHTMICCRDEVN	0.8466
TEBD4	MCSCKLLVVALFAVFLISLGSS	LNLSARVCWQKGGRRCHRAGQCYDNEIEIGTCYHLRMKCCRDKSQEETPT	0.4649
TEBD5	MRFLLLTFALLLFYVASG	QLFECMKDGGFCQTGLFKECEYGLPYNCGINAICCKRGPVRSISFLVMR	0.8693
TEBD6	MSSCKLLVVALFAVFLISLGSS	LNLSARVCWQKGGRRCHRAGQCYDNEVEIGMCMYHLRLKCCRDKSQEETPT	0.4446
TEBD7	MRFFLLTFALLLFYVASG	KIMNCKSDGGFCQTGLFKECEYGLPYNCGINAICCKRGPVRSISFLVMR	0.8546
TEBD8	MRFLFITFALLLFYVSSG	QPEACKENGGFCQIGLQKECVYGLPYNCGINAKCCKRSWVRSISFLVGR	0.5772
TEBD9	MKICYLLGLIALLAYLPLPGATQG	SNACRIKRGYCYFLSCKPQTRQIGRCTNRHPCCRW	0.606
TEBD10	MKICYFLLGIALPAYLPLPGAMQ	GIQQCLKGGFCRVSACPGGTIQIGHCEPVRLLCCKR	0.254
TEBD11	MNILYCFYAVIFLFFHAAQEDQFIEA	ILLTCASVGGFCILQPNETCPVSGVLLDVPCHFGRRCCSKTDV	0.2194
TEBD12	MKLYHLFFLTLFVKTVVG	PDIAYDCLTDFDGFYPPDLPCPGCFIPYGNCFKLQCCAK	0.3797
TEBD13	MKPWFVLLFFTFLLSDLGEA	DMIKESRSICKAKGTCRKGSKSSEKKIGLCHVNQPCCKS	0.9932
TEBD14	MKPWFVLLFFTFLLSDLGEA	DMIKESRSICKAKGTCRKGSKSSEKKIGLCHVNQPCCKS	0.9932
TEBD15	MKLWCVLLFLTFAILSDDLGEA	ENGKECIAKNGFCHAQCPVSNYKKNCKSDVACCLM	0.9673
TEBD16	MKTLFLLFAALLFFSQIIPG	SFQPAPDTLQCRSSHGFCKAYYCPHTIPTGGSCQWGSLLICCKS	0.2872
TEBD17	MKAFLLLVAIFMLSYQAVTVTG	QRDPANIACFQSGGTCRASCFFPSVQSGDCAGGFVCTW	0.7616
TEBD18	MKALLLLVAIFLFSHQAAAS	DPNDPRDIACRRIGGYCVWEYCPYTTFYNGPCSDCKACCTW	0.8499
TEBD19	MKALLLLVAIFLFSHQAAAS	DPNDPLDIACRKGGSCEWRKCPPTIFYNGPCSGGMACCYW	0.8576
TEBD20	MKILYLLFAFLFLVFLSEPGNA	QSKCRRERIGICYGRVGVSTSDIGRQDCGPRSRCCQR	0.9459
TEBD21	MKALLLLVAIFMFSHQAAAS	DPNDPRDIACKMRGSCSEWRRCPPPTTVTRGACSGRMACCWSQVNLVLLFCSKQMQCCFFIYTIYF	0.8255
TEBD22	MKILYLLFAFLFLVFLSEPGNA	QSKCYHKGGGCAYGHCPDSTLDIGRQDCGPRSKCCRG	0.9394
TEBD23	MKALLLLVAIFMFSHQAAAT	VDFNFAETDCPIDVGFCLDSCDYLGTPYRCPYGGICCLW	0.665
TEBD24	MMTYLLFAFLFLVFLSESGNA	QRWCHRQGGRCFSHRCLQNFENLGKIDCRQSHVCCRP	0.926
TEBD25	MKTLFLLFAALLFFSQIIPG	SFQPAPDTLQCRSSHGFCKAYYCPHTIPTGGSCQWGSLLICCKS	0.2872
TEBD26	MKAFLLLVAIFMLSYQAVTVTG	QRDPANIACFQSGGTCRASCFFPSVQSGDCAGGFVCTW	0.7616
TEBD27	MKALLLLVAIFLFSHQAAAS	DPNDPRDIACRRIGGYCVWEYCPYTTFYNGPCSDCKACCTW	0.8499

TEBD28	MKILYLLFAFLFLAFLSEPGNA	QRKCRERGRRCYGRVGFLLDIGRQDCRWRARCCRR	0.9807
TEBD29	MKALLLLVAIFLFSHQAAAS	DPNDPRDIACRKIGGSCEWRKCPPTIFYNGPCSGGMACCYW	0.8525
TEBD30	MKALLLLVAIFLFSHQAAAS	DPNDPRDIACKKMRGSCEWRRCPPTTVTRGACSGRMACCSW	0.859
TEBD31	MKILYLLFAFLFLAFLSEPGNA	QSKCYHKGGGCAYGHCPDSTLDIGRQDCGPRSKCCQG	0.9398
TEBD32	MKALLLLVAIFMFSHQAAAT	VDFNFAETDCPIDVGFCDSCDYLGTPYRCYPYGGICCLW	0.6638
TEBD33	MKALLLLVAIFMFSHQAAA	IDDNIIIDCPINVGACLVDCYHLLSPYRCPGQICCCQW	0.9534
TEBD34	MKIFYLLVLAFLFFAVLPESGYA	LYLCYSRGGHCVPANSCTPERDLGTWGCNTGLTCCRR	0.9289
TEBD35	MKTLFLLFAALLFFSQIISG	SSQTAPDTLECRSSHGHGFKSRCPPTIPTGGSCQWGLICCKS	0.4202
TEBD36	MKALLLLVAIFMFSNQAVTA	TGQSDPANIAFCQSGGTCRASCPFGVQSGDCAGGFVCTTW	0.588
TEBD37	MKALLLLVAIFIFSHQAAAT	DDPLEVACTNKGGSWSKCPYPSVNAGRCRYPQVCCCTW	0.7122
TEBD38	MKIFYLLFAFLFLAFLPEPGNA	GSECRSHGGVCDNSCGSGYYPYIGKYDCGTGWCCAP	0.9469
TEBD39	MKIFYLLFAFLFLAFLPEPGNA	GSQCRSHGGVCDNSCGRGYYSIGKYDCGTGWCCAP	0.9602
TEBD40	MKIFYLLFAFLFLAFLLEPGNA	DRLCFNAGGKCFPQCPDGYTSIGNCNKIGICCKK	0.9927
TEBD41	MKIFYLLVLAFLFLAILPEPGNA	NYVCRRRSGDCVPSNSCPPERNLGTWGCNTGLTCCRR	0.9809
TEBD42	MKIFYLLVLAFLFLAILPEPGYA	HYVCFRQGGVVCVHSCPPGRNLGTWGCNNRLTCCRR	0.9881
TEBD43	MMTLYLLFAFLFLAFLSESGNA	QRWCHRRGRCFSHHCLQNFENLGKIDCRQSHVCCRP	0.9236
TEBD44	MKILYLLMFAFLFLAFLSEPGNA	QRRCHRQGRSCFRRCPRRYKNLGRWNCRRRFTCCQL	0.9768
TEBD45	MKIFYLLFAFLFLAFLSEPGNA	QCRSRGGGLCYRRHCPINTVSFGRLDPCWILRCCVP	0.9287
TEBD46	MKIFYLLFAFFIFLSEPGYA	QRKCHHKNGVCVNPCHKSAIPINIGEDCKDRATCCRP	0.9868
TEBD47	MKIFYLLFAFLFLVFLSETGNA	QRQCHRAKGSFPRPCPRKSNLGVDCSGRMICCKP	0.9519
TEBD48	MKILYLLFAVLILAFLSQPGNA	QDRCHSIRGRCPNRCPGGLDHGQVDCRYRWRCCVR	0.9857
TEBD49	MKILYLLFAVLFLAFLSQPGNA	HHLCARRGGICRRRCRRGNEQSHGQVDCQLGLQCCVR	0.9849
TEBD50	MKIIYLLFAVLFLAFLSQPGNA	QPRCRGLGGICRPGRCRPGQHCFGQIDCRRGWKCCRR	0.9797
TEBD51	MKIIYLLFAVLFLAFLSQPGNA	QRRCRNRGGYCCRTRCPRPHHCFGRMDCPPRHNCRR	0.9822

*Signal peptide cleavage sites on the Beta-defensin showing signal sequence and mature peptide sequence. The probability shows the likelihood that the cleavage site is in that position.*

### A3.1 *Chelonia mydas* exon positions

GENE	Exon 1		Length (bp)	Exon 2		Length (bp)	Orientation	Total Length (bp)	No of AA
	FROM	END		FROM	END				
CMBD1	59156	59213	58	56361	56482	122	-	180	60
CMBD2	70952	71009	58	69075	69202	128	-	186	62
CMBD3	86628	86685	58	85687	85820	134	-	192	64
CMBD4	96547	96607	61	98541	98683	143	+	204	68
CMBD5	115369	115426	58	116254	116387	134	+	192	64
CMBD6	130514	130571	58	131383	131534	152	+	210	70
CMBD7	143871	143928	58	141967	142100	134	-	192	64
CMBD8	152849	152906	58	152034	152167	134	-	192	64
CMBD9	167474	167531	58	165775	165899	125	-	183	61
CMBD10	188114	188171	58	184575	184708	134	-	192	64
CMBD11	197873	197930	58	193284	193417	134	-	192	64
CMBD12	199695	199752	58	199192	199322	131	-	189	63
CMBD13	208908	208965	58	207984	208111	128	-	186	62
CMBD14	249114	249261	148	248284	248417	134	-	282	94
CMBD15	267912	268065	154	269867	269994	128	+	282	94
CMBD16	279643	279697	55	281599	281726	128	+	183	61
CMBD17	288312	288366	55	290045	290172	128	+	183	61
CMBD18	297856	298009	154	299813	299940	128	+	282	94
CMBD19	312292	312451	160	309801	309928	128	-	288	96
CMBD20	323408	323555	148	322169	322299	131	-	279	93
CMBD21	339780	339927	148	338544	338671	128	-	276	92
CMBD22	376121	376178	58	374511	374650	140	-	198	66
CMBD23	384201	384258	58	386108	386247	140	+	198	66
CMBD24	428205	428262	58	431875	432041	167	+	225	75
CMBD25	448240	448294	55	446645	446769	125	-	180	60
CMBD26	461627	461684	58	459075	459238	164	-	222	74
CMBD27	521992	522049	58	516623	516789	167	-	225	75
CMBD28	564502	564559	58	562761	562912	152	-	210	70
CMBD29	650500	650557	58	652228	652352	125	+	183	61
CMBD30	803581	803638	58	805865	805992	128	+	186	62
CMBD31	839154	839211	58	837649	837776	128	-	186	62
CMBD32	856588	856645	58	854315	854442	128	-	186	62
CMBD33	865051	865108	58	867262	867392	131	+	189	63
CMBD34	910920	910977	58	909544	909668	125	-	183	61
CMBD35	928840	928897	58	927350	927474	125	-	183	61
CMBD36	972334	972391	58	973725	973849	125	+	183	61
CMBD37	984577	984634	58	986026	986153	128	+	186	62
CMBD38	1001087	1001144	58	1003213	1003340	128	+	186	62
CMBD39	1027124	1027178	55	1030804	1030940	137	+	192	64

Positions coordinates of exons. Last codon of CTSB is a marker for the start of the beta-defensin cluster. Orientation of the gene and length of exons and full length of coding sequence.



### A3.2 *Chelonia mydas* physical properties of second exon mature peptide

	pI	Net Charge	Mr		pI	Net Charge	Mr
CMBD1	8.69	3	4549	CMBD21	3.93	-7	7587
CMBD2	8.34	4	4687	CMBD22	7.79	1	4903
CMBD3	6.27	-1	5093	CMBD23	7.83	1	4703
CMBD4	9.69	8	5777	CMBD24	9.73	9	5990
CMBD5	4.63	-4	5065	CMBD25	8.29	2	4494
CMBD6	4.63	-6	5755	CMBD26	9.49	7	6075
CMBD7	8.98	4	4904	CMBD27	9.18	6	6117
CMBD8	6.24	0	4929	CMBD28	8.53	3	5472
CMBD9	5.92	0	4336	CMBD29	7.8	1	4531
CMBD10	8.33	2	5028	CMBD30	9.69	6	4811
CMBD11	8.98	4	4577	CMBD31	8.31	2	4931
CMBD12	9.48	6	4879	CMBD32	9.8	7	4968
CMBD13	9.7	8	4843	CMBD33	10.63	8	4932
CMBD14	6.12	-1	8444	CMBD34	9.61	7	4883
CMBD15	7.75	1	7941	CMBD35	9.38	6	4736
CMBD16	10.6	8	4963	CMBD36	9.7	8	4610
CMBD17	10.2	8	4866	CMBD37	8.7	5	4533
CMBD18	5.73	-1	8004	CMBD38	9.7	8	4806
CMBD19	4.54	-5	8424	CMBD39	8.66	3	5179
CMBD20	4.02	-7	7604				

*Physical properties of the Chelonia mydas mature peptide beta-defensins. All properties were achieved by using the protparam program on the ExPASy Server (Gasteiger, E. et al. 2005)*

### A 3.3 *Chelonia mydas* signal peptide prediction

	Signal Peptide	Pro-peptide/Mature Peptide	Probability
CMBD1	MRILYLFAVVIFFLQAAPTRG	SAYDTLQCLSNHGHCRPLCFHMERQVGTCTNGHQRCCK	0.5315
CMBD2	MKIFYLLFAGLFLVLPNPGNG	QFVILGCLTRGGSCRTDNCYLDEMEIGSCLRSNRLCCKRT	0.9264
CMBD3	MRILLLLLVVLVSGISLATSANG	QRVTRYLNHCLQRGGTCRYDDCDDGEVQIGTCYHHTMVCCR	0.4984
CMBD4	MSILNLVLPVFLMLLVEAPG	DWKIIPFLDSLTCGLKHYRCRKTFCYLNERKLGMCILRSRFFCRRMT	0.4133
CMBD5	MRILYLLVVLVSGISLATSANG	QRMTRDLSYCLERGGTCQYNDSCSDGEVEIGTCYHHTMLCCWD	0.5237
CMBD6	MRILYLLVVLVSGISLATSANG	QRMTRDLSHCLQRGGTCRYNDCCDGEVEIGTCYHHTMLCCQDEEEMLS	0.2129
CMBD7	MKILCLIFAVLLFLLQATPGLS	PPSDTLRCISNNGLCHRTLCPRLFAFGICSHGRETCCKGRW	0.2216
CMBD8	MRSLYLLFAVGVLLFQAPGYG	QMREIAACVSDGYCVEAFKMCPSGEYLLGICPDFIMRCCCK	0.7805
CMBD9	MRSLYLLFVVALLFHAAPGDG	MPFNTCQSRRGFCLAALQICPSGIFLDLLCIGGGGCCQI	0.7413
CMBD10	MRSYLLIAVVPFLFQAPGYG	QRMQIPCVSRGGYCVFAEFCPSHEYLRGACPNFIMRCCCK	0.8952
CMBD11	MRGLYLLFAVVLFLFAAPGSG	LRLPFVPCLLKGGCLPRLGLCSPGIQLLRGVCPAPLICCCQT	0.6916
CMBD12	MKILYLLFAVVLFLFQAAAPGSG	EEARIPSCRFMGGYCIARRGQHCPSGRFLNGPCGFRERCCCKR	0.8036
CMBD13	MRILYLLVLLVFLGLQTA	QDTRKCKSNWYHMGGRCLWGCKFGEKRTGTCLAGLMTCCHPKH	0.4001
CMBD14	MKILYLLFAVVLVFLVHVQG	QHQQEPQDDPQAWNEAPDDAEAEAEVAPGPKQOSHILCSFRGVCRSKRCSKQERKIGNCAYRACCVKRN	0.9677
CMBD15	MKILYLLFAVVLVFLVQQTSEA	VALTKNEAEAQAPDAVEAEAAQDAAEISSPDLMPQKAPIFCAMILGVCRIRCSLRETRIGWCSRGVSCCRKF	0.6825
CMBD16	MKILYLLFAVVLVFLVQQTQN	LMPQEAPIHARQLGVCRIRCSQRERRIGWCSRGMACCRKRF	0.4012
CMBD17	MKILYLLFAVVLVFLVQQTQN	LMPQEAPIFCAMKLGVCRIIRCSRRERTIGSCFRGVSCCRKRF	0.4055
CMBD18	MKILYLLFAVVLVFLVQQTSEA	VAVIKNEAKAQAPDAVEAEAAQDDAEISSPDLMPQEAPIHARQLGVQCTKCSWRERTIGWCSRGVACCKKWF	0.7393
CMBD19	MKTYLLFAVACLVFHVQA	SPKLPEDVVPQDESENLDAAEDDGIGMEDGDIADKADPDVQVSPFQCLRRARGVCRPLRCNKNEGTRICFCYRVPCCSK	0.9265
CMBD20	MKTYLLFAVCLVSHVQA	NPMVPVEDVVPQGEQNLDDGIGVEDVDVVEAPGGQSNPMVCSFSGGTCCKGSDSGKEVTSGMICYPGVMCCIRKP	0.9738
CMBD21	MKTYLLFAVCLVSHVQA	NPMVPVEDVVPQGEQNLDDGIGVEDVDVVEAPGGQSNPMVCSFSGGTCRNYGCVGREVTSGMICYPGVCCIRP	0.9391
CMBD22	MRILYLLFAVVLVFLVQGPAGNA	DFLDNINCRSNFQFCHAGDCPISTTLVGTICINGKINCCKRPTAP	0.9162
CMBD23	MKILYLLFAVVLVFLVQATPGAA	DMGPPPADTLACRAQGGFCHLNCPPVFSISGTCHGGQLKCCTR	0.804
CMBD24	MKIVYLLFAVFLVLQSTPGFT	QFINSALCKRARGSCRRICYGKYRLIGSCGSGQNCCMRAVSSGCHKGDITV	0.8121
CMBD25	MKILYLLFAVVLVFLVQVAPGLA	QNIWTPSQCGCFGGDCHVPCPHGTRHFGRCVTQGFCCLR	0.9142
CMBD26	MKILYLLSVLFFVLQIIPGFTYS	LGATMRCLQNGGRCPWQCPANTYNIGRCCSWRLCCRRVSSGLHRGNGTV	0.3885
CMBD27	MKIVYLLFAVFLVLQSTPGFT	KFISNPFACVRRAGGFCTYCYKYGWFGTCGSGQTCRRRWVSSGCHKGDITV	0.7184
CMBD28	MKILYLLFALLVFLVQSSPGFT	DPRQCLGRGEFCRTRCCSPSIQIGVCAIGIPCCKDRVSSQCHKADIIV	0.6704
CMBD29	MKILYLLFAVFLVFLVQSSPGFT	QRPPSDPESCRRAAGGVCHFICSPFTFPFGTCGFVESCCR	0.8185
CMBD30	MKILYLLFAVFLVLAQS	TEVSNRGIIGTISICLSRRGACFLFHCPLNTRIGRCGLFWHCCR	0.6007
CMBD31	MKILYLLFAVFLVFLVQGAPEFSQA	ESPYLQCRHLGGDCYLKRCHEFNFAIGICDKYQVCCCKR	0.7429
CMBD32	MKILYLLFAVFLVFLVQDAPEFSEA	QDPFILCRSRRGFCYSYKRCPFNSTLISRCSGRFLCCRR	0.9034
CMBD33	MKILYLLFAVFLVFLVQGAPEFSQA	QNFNRRCILRGGTCFYLRCPFLRTRIGRCFSGGVCCAR	0.4041
CMBD34	MKFLYLLSAIVFLVMDAPGFS	HGLLTHHACNRNGGHCHFWKCPYHTRYLGKCVFGHCCRR	0.4962
CMBD35	MKFLYLLSAIVFLMLIDAPGFSHA	LLTHKACRRRAGDCHFWKCPYHTRYLGKCVFGHCCQR	0.5052
CMBD36	MKILYLLSAAVFLVLLTAPGFSQA	KISPNKCKRRRGSCYFRGCPFSSVYLKCVWIGSCCQR	0.6937
CMBD37	MKILYLLFTVFLVFLVQGVSVFS	QAERSSACGELGGACFLRCPNSVHVRSCYPQGVCCRR	0.442
CMBD38	MKILYLLFALVFLVFLVQGAPEFSQA	WRSRKRRCGRVAGICTGPCPYNYILIGICSRKYSCCKL	0.5333
CMBD39	MKVLYLLFLVLYFFQGTSG	TGRCRRLNGVCRHTLCHHVETVYVGRCHHGMGNCLNDDDDRKLKV	0.7337

Signal peptide cleavage sites on the Beta-defensin showing signal sequence and mature peptide sequence. The probability shows the likelihood that the cleavage site is in that position.

### A3.4 *Gopherus evgoodei* exon positions

GENE	Exon 1		Length (bp)	Exon 2		Length (bp)	Orientation	Total Length (bp)	No of AA
	FROM	END		FROM	END				
GEBD1	69307	69364	58	61924	62045	122	-	180	60
GEBD2	80172	80229	58	77979	78112	134	-	192	64
GEBD3	90732	90789	58	89847	89983	137	-	195	65
GEBD4	97766	97823	58	103371	103513	143	+	201	67
GEBD5	111242	111299	58	114074	114207	134	+	192	64
GEBD6	134301	134358	58	135174	135307	134	+	192	64
GEBD7	146152	146209	58	141871	142004	134	-	192	64
GEBD8	151365	151422	58	150551	150684	134	-	192	64
GEBD9	161548	161605	58	159912	160036	125	-	183	61
GEBD10	194757	194814	58	190269	190402	134	-	192	64
GEBD11	204799	204856	58	200227	200351	125	-	183	61
GEBD12	207824	207881	58	206095	206225	131	-	189	63
GEBD13	245275	245329	55	246207	246334	128	+	183	61
GEBD14	271335	271389	55	273902	274026	125	+	180	60
GEBD15	287500	287629	130	289001	289116	116	+	246	82
GEBD16	305434	305491	58	303830	303969	140	-	198	66
GEBD17	310269	310326	58	313075	313214	140	+	198	66
GEBD18	343397	343451	55	341335	341459	125	-	180	60
GEBD19	358136	358193	58	355619	355782	164	-	222	74
GEBD20	393134	393191	58	391889	392031	143	-	201	67
GEBD21	415976	416033	58	413543	413685	143	-	201	67
GEBD22	434359	434416	58	432709	432875	167	-	225	75
GEBD23	463296	463353	58	460879	461072	194	-	252	84
GEBD24	533075	533132	58	529947	530128	182	-	240	80
GEBD25	550091	550148	58	547977	548143	167	-	225	75
GEBD26	561586	561643	58	559824	559954	131	-	189	63
GEBD27	621332	621389	58	619259	619389	131	-	189	63
GEBD28	702925	702982	58	700354	700520	167	-	225	75
GEBD29	773141	773198	58	774488	774618	131	+	189	63
GEBD30	838829	838889	61	847500	847633	134	+	195	65
GEBD31	911268	911325	58	913526	913689	164	+	222	74
GEBD32	937415	937472	58	939798	939925	128	+	186	62
GEBD33	1008420	1008459	40	999640	999767	128	-	168	56
GEBD34	1023593	1023650	58	1022117	1022241	125	-	183	61
GEBD35	1057120	1057177	58	1054788	1054915	128	-	186	62
GEBD36	1084550	1084607	58	1086283	1086410	128	+	186	62
GEBD37	1099989	1100046	58	1102310	1102440	131	+	189	63

GEBD38	1125792	1125849	58	1127251	1127408	158	+	216	72
GEBD39	1158282	1158339	58	1155657	1155781	125	-	183	61
GEBD40	1172438	1172495	58	1170961	1171085	125	-	183	61
GEBD41	1183174	1183231	58	1181696	1181823	128	-	186	62
GEBD42	1202506	1202563	58	1201028	1201155	128	-	186	62
GEBD43	1232328	1232385	58	1237167	1237294	128	+	186	62
GEBD44	1262804	1262861	58	1264151	1264275	125	+	183	61
GEBD45	1273453	1273510	58	1275241	1275356	116	+	174	58
GEBD46	1290798	1290855	58	1295901	1296028	128	+	186	62
GEBD47	1319964	1320018	55	1325822	1325958	137	+	192	64

*Positions coordinates of exons. Last codon of CTSB is a marker for the start of the beta-defensin cluster. Orientation of the gene and length of exons and full length of coding sequence*

A3.5 *Gopherus evgoodei* Physical properties of mature peptide

	pI	Net Charge	Mr		pI	Net Charge	Mr
GEBD1	9.18	5	4625	GEBD24	6.87	0	6634
GEBD2	8.35	2	4565	GEBD25	8.75	4	5934
GEBD3	6.42	-1	5099	GEBD26	5.44	-1	4614
GEBD4	9.18	5	5514	GEBD27	8.53	3	4661
GEBD5	6.88	0	5153	GEBD28	9.08	5	6179
GEBD6	4.97	-4	5037	GEBD29	5.27	-2	4629
GEBD7	9.22	5	4841	GEBD30	10.79	10	5028
GEBD8	4.94	-1	4874	GEBD31	11.1	9	6108
GEBD9	8.69	3	4392	GEBD32	9.22	5	4871
GEBD10	7.79	1	4846	GEBD33	9.38	6	4986
GEBD11	9.56	6	4343	GEBD34	9.12	5	4761
GEBD12	9.99	9	4889	GEBD35	9.5	7	4772
GEBD13	9.99	8	4724	GEBD36	9.69	7	4642
GEBD14	9.55	7	4833	GEBD37	9.69	7	4928
GEBD15	9.38	6	4121	GEBD38	10.63	8	6123
GEBD16	6.7	0	4978	GEBD39	9.55	7	4908
GEBD17	6.68	0	4775	GEBD40	9.38	6	4908
GEBD18	8.87	4	4549	GEBD41	8.9	4	4648
GEBD19	9.04	6	5856	GEBD42	8.87	4	4912
GEBD20	9.77	9	5519	GEBD43	8.66	3	4900
GEBD21	10.34	9	5413	GEBD44	9.7	8	4705
GEBD22	9.84	8	6326	GEBD45	7.68	1	4572
GEBD23	10.36	11	7399	GEBD46	10	9	4661
				GEBD47	8.3	2	5241

*Physical properties of the Gopherus evgoodei* mature peptide beta-defensins. All properties were achieved by using the protparam program on the ExPASy Server (Gasteiger, E. et al. 2005)

### A3.6 *Gopherus evgoodei* Signal peptide prediction

	Signal Peptide	Mature Peptide	Probability
GEBD1	MRILYLFFAFVIFFLQAAP	ARGSAYDTLQCLSKHGHCRRLCFHMERQVGTCTNGHMRCCK	0.6375
GEBD2	MKIFYLLFACLFLVSLPSPGNG	QFAILGCLIRGGSCRTDNCSLDEMEIGSCLRSNRICCKRA	0.8482
GEBD3	MRILCLLLVLVLLGISLFTSANG	QRMTRHLNHCLQRRGGTCRYDDCDDGEVQIGTCYHHTMVCCR	0.8733
GEBD4	MSILHVLFPVFLVLLLAGPGDG	NLLHLIDSLACKGKHGRCREAFCLNERQIGMCTFHTRFCCRRQK	0.965
GEBD5	MRFFYCLFAVLLWIALA	APGKGVNHGAVSCRNRDAVCQFDSCYYNEIEIATCYHYTMKCCRERD	0.4346
GEBD6	MRILYLLLVLVLLGISLFTSANG	QRMTRDLSHCLERGGTCRYNDCDDGEVEIGTCYHHTMLCCQD	0.8712
GEBD7	MKIFYLIFAVLLFHLQAAPGLS	LSADTLRCISNNGLC HQTLCPRTPFKFGTCSHGRATCCKGRW	0.287
GEBD8	MRSLYLLFAVGVLLFQLAPGYG	QMTEIPACVSVGDYCVFAFKMCPGSEYLLGICADDFIMRCCCK	0.9206
GEBD9	MRGLYLLCVVALLLFNAGPGDG	IPIYSCQSSRRGFCLLDFRLCPSGIALALACSPGRCCKI	0.9231
GEBD10	MRSLYLLIAVLLVLLFQAPGNG	QRIQMPPCASIGGYCMEALEICPSHEYLRGVCAHFLMGCCCK	0.9471
GEBD11	MRGLYLLFAVLLVLLHA	APESGLRPLFAPCLVKGGLCRPRFLCSAGIQLLRGVCPAPLRCC	0.2659
GEBD12	MRGLYLLFAVLLVLLHAAPGSA	YKARI PSCRSLRGYCIIRRGQSCHSGQYLNACPPRERCCCKR	0.847
GEBD13	MKILLLLFAVVFLVFQTQS	LRGRGGVVACQNRGICRGFCWLNEYASGRCTGKLCNRKN	0.6897
GEBD14	MKILLLLFAVVFLVFQTQNRG	RFLRPTMCQLGGVCRGRCFSYEYPSAWCFNMYCCKRK	0.6036
GEBD15	MQALYLLFALLFLVFQEQAQS	KEQDEPQEPALLDQIEGARIVREISSSCLARGGQCRLGFCPWKETKITSCGFGRPCCKKVI	0.9398
GEBD16	MRILYLLFAVLLFQAGPNA	DFLDNLNCRNNFGFCHSGDCPPSTTLIGTCINGKINCKWTTAP	0.9188
GEBD17	MRILYLLFAVLVFLFQGTGVA	DLGPPLADTMACRDQGGFCQLMSCPQVFSVSGTCHGGLLKCCTR	0.8251
GEBD18	MKILYLLFTVLLVLIAPDLA	KTIWTPSQCEHFGGVCSAPCPRCTRQFGRCVTQGFCCLR	0.4794
GEBD19	MKILYLLFAVLFLVLQSSIPG	FTCCPGATIRCLQNGGRCPYQCPPNTYIIGHCCPWRLCCRRVSSGLHKGNGTV	0.4759
GEBD20	MKILYLLFAVLFLVLQSTPGLT	QSPTCFMRCIRRGGLCYRRCPPGTYYLGRCCRQFFCCRRVSAGFQ	0.8416
GEBD21	MKILYLLFAVFLVLQSSSGLA	QYINSDAVCSRLGGRCFPICYSPIWIKIGNCRFRSRRRRVSSG	0.7358
GEBD22	MKILYLLFALFFLVLQSSPGFT	QFINNPFACRRRARGICRRSCYPNLRPIGRCGFAQSCRRSWVSSGCHKEDITV	0.7901
GEBD23	MKILYLLFALFFLVLQSSPGFT	QFINNSFACRRRARGSCRRFCIGRYRLIGTCGQGNCCRRRVSSGCHKRVITVQTEYLCLSFL	0.7647
GEBD24	MKILYLLFAVFLMLQSTTG	ITDPRQCIGHGEFCSIRCHPPSRQIGICAIGIPCCKRQVSSGFNEADVTVYTEYLCLSFL	0.7296
GEBD25	MKILYLLFAVFLVLQSSPGFT	RYPACEVRKCIQNGGLCFGTCPAPFRETGSCGCGVSCCQWRVSSGFHKANVTL	0.6721
GEBD26	MKILYLLFAVFLVLQSSPGFT	QCPADNVHECIQNGGLCFSTCPAPFRETGSCGCGVSCCQWR	0.7287
GEBD27	MKILYLLFAVFLVLQSSPGFT	RYPACEVRKCIQNGGLCFGTCPFPFIQSGSCGCGVSCCQWR	0.6644
GEBD28	MKILFLLFALFFLMLQSSPGFT	HFINDPEACRRAGGFCLRRCAPYFTPIGSCGIVQSCRRRVSSGCHKTDVIV	0.7275
GEBD29	MKIFYVLFVAVFLVLQSSPGFA	QDFPSYPEACIHAGGFCHFSCPLLSTPIGSCGFVESCRRWG	0.9656
GEBD30	MKIFYLLFALFFLVAQSGA	QESDRNIGAIIPVAVCIARKGKCYFRRCPPRRTRIGRAVFFPCCR	0.8952
GEBD31	MKILYLLFAVFLVAQS	TEVSNRGI IATARCLRRRGACFLFNCPITYTVRIGRCGVFWHCCRRVSPGLHTGDIAR	0.6025

GEBD32	MKILYLLFAVFFLVAQS	TEVADRGIFGTAMCVSRKGACFLFHCPLYTMRIGRCGLFWHCCRW	0.5663
GEBD33	MRFFGFLVLLLSSEFSQA	KDPFLQCRLRGGNCYFKRCLFNSKILGICDRLHFCCQR	0.9861
GEBD34	MKILFLLFAVVFLVLM DAPGFSQA	KMSRRECEHRGGECYPMCPKFYKTIGNCVDGGRCCRR	0.4592
GEBD35	MKILYLLFAVFFLV LQGMPEFSEA	QDPFKLCKYRGGTCSYKRCSFNSKVVGICGGRFLCCRR	0.9017
GEBD36	MKILYLLFAVFLV LHCVPAFSQA	YDLNRNCRLRGGTCYIGKCPRAVRSGICSRGNVCCLT	0.5333
GEBD37	MKILYLLFAVFLV HQGATEFSQA	QNLNRRCVLLGGTCFYRRCPPFIRTVLGRCYKGDVCCGR	0.4097
GEBD38	MKILYLLFAVFFLV LQGTPAFS	QAQNSSFLCRRLWGTCLRRCPPNWRFIGRCSSTHVCCVRYVQLSTRQTL	0.5186
GEBD39	MKILYLLFAVVFLV LMDAPGFS	YALLTRRGCRKGGECNFWRCPSNAIYLDKCYFGHCCRR	0.5148
GEBD40	MKILYLLFAVVFLV LMDAPGFS	YALLTQRGCRRRGGECNFWRCPSNAIYLDKCYFGHCCRR	0.5139
GEBD41	MKIFYLLFAVVFLV LIDAPGFSQA	LVTGGCKRYGGSCFFWRCPTPSTYFDKCI PVGHCCVK	0.4732
GEBD42	MKILYLLFAVAFLV LMDAPGFS	QAWVNSRNCKYRGGSCYFWQCPTTSTYIDKCI PFGYCCVK	0.5302
GEBD43	MKNLYQIFAVVFLV LMDVPAFSQA	QLTRWQCRWHGGDCYNPVCPIYISKLIGRCIPFGYCCQT	0.4838
GEBD44	MKILYLLSAVVFLV LLTAPGFSQA	RISPKECKRRGGSCYFRGCPSNSIYLKCKWIGSCCQR	0.6655
GEBD45	MKILYLLFAVVFLV LQGVPG	SEDQLGVYMAWRCLLAFEMSGQSMYIRRFYPQGFCCQR	0.6095
GEBD46	MKILFLLHAVLFLV LQLASEFSQA	WRSSKRCRRAGGFCSGPCPSNAKLIGICSRKYSCCKL	0.7524
GEBD47	MKVLYLFLVFFYFFQGTSG	TGRCRRLNGVCRHTLCHHIETYVGRCHHGMGNCLNDDDDRKEKM	0.7286

*Signal peptide cleavage sites on the Beta-defensin showing signal sequence and mature peptide sequence. The probability shows the likelihood that the cleavage site is in that position.*

A4.1 *Alligator mississippiensis* exon positions

	GENE	Exon 1		Length	Exon 2		Length	Orientation	Total Length	No of AA
		FROM	END	(bp)	FROM	END	(bp)		(bp)	
>NW_017709158.1:1-259212 rev comp	AMBD1	51633	51693	61	49208	49329	122	-	183	61
	AMBD2	62922	62964	43	61912	62078	167	-	210	70
	AMBD3	66092	66152	61	65079	65218	140	-	201	67
	AMBD4	102333	102390	58	108229	108350	122	+	180	60
	AMBD5	129140	129197	58	130946	131070	125	+	183	61
	AMBD6	152828	152888	61	156333	156469	137	+	198	66
	AMBD7	180419	180476	58	178930	179051	122	-	180	60
	AMBD8	195370	195526	157	193802	193971	170	-	327	109
	AMBD9	206355	206520	166	204668	204798	131	-	297	99
	AMBD10	228087	228258	172	226940	227058	119	-	291	97
	AMBD11	240552	240723	172	243006	243133	128	+	300	100
	AMBD12	252138	252306	169	253666	253793	128	+	297	99
	AMBD13	265719	265848	130	267784	267917	134	+	264	88
>NW_017710918.1:222936-2-2365424 rev comp	AMBD14	14960	15119	160	14447	14571	125	-	285	95
	AMBD15	59012	59069	58	60730	60857	128	+	186	62
	AMBD16	70239	70296	58	69314	69495	182	-	240	80
	AMBD17	89590	89647	58	87981	88096	116	-	174	58
	AMBD18	98923	98980	58	101911	102044	134	+	192	64

Positions coordinates of exons. Last codon of *CTSB* is a marker for the start of the beta-defensin cluster. Orientation of the gene and length of exons and full length of coding sequence.



#### A4.2 *Alligator mississippiensis* physical properties of mature peptide

GENE	pI	Net Charge	Mr
AMBD1	9.18	5	4576
AMBD2	8.9	4	5336
AMBD3	5.13	-2	5106
AMBD4	8.38	2	4195
AMBD5	7.78	1	4064
AMBD6	7.79	1	4716
AMBD7	9.98	7	4291
AMBD8	5.61	-3	10292
AMBD9	4.97	-4	8957
AMBD10	4.38	-8	9025
AMBD11	4.38	-8	9025
AMBD12	9.58	8	9399
AMBD13	5.74	-1	7367
AMBD14	5.17	-3	8105
AMBD15	8.84	4	4920
AMBD16	8.72	3	7055
AMBD17	8.58	3	4241
AMBD18	9.1	5	5155

*Physical properties of the Alligator mississippiensis mature peptide beta-defensins. All properties were achieved by using the protparam program on the ExPASy Server (Gasteiger, E. et al. 2005)*

### A4.3 *Alligator mississippiensis* signal peptide prediction

	SIGNAL PEPTIDE	Pro-domain/MATURE PEPTIDE	Probability
AMBD1	MRILYLLFAAVMILFLQAVPAKG	SYSTLQCRNNHGHCRRLCFHRERWIGNCNGGHQHCK	
AMBD2	MLWFAVAILLAVPGNAQG	SKNVCRSAGGQCMGTCLSSEVIGDCFTPVILCKKYLARKTPGELQGGGA	0.5155
AMBD3	MMKFFYLLLVFLGIFLATTANG	QRASRYVNHCLQKGGTCRYDDCEAGEEQIGTCYRQTMVCCRDEE	0.9039
AMBD4	MKNLYLILALFFSQVAPGGA	APSPHEICRRHGGTCVISISFCTHLIVEVLGCICCRQR	0.6614
AMBD5	MKSLYVILAVALFFSQVVPNGG	LPILSLIQCLNLGGICLISVSLCDGVTIRLLGCNCCSSR	0.6013
AMBD6	MRVLCLLLIVITLLFQAAPGYS	QRTISPLCDSVGGYCVNPFVCLSGREIVGSCPHLLMRCCCKMI	0.8231
AMBD7	MRILYLLFAVFLQVAPGQS	YRECRNRGGECRPHGSGHPGVSIPVRCPHRTVCCRRR	0.7088
AMBD8	MKLLFLLGVTTLVFQAQA	QDVVVAQDEAEPQDLGEMEEEAETEVM EAEDATGPKLGESPAHCRWKRGVCRRTTHCKRNDRNCRHTPCPAERIIGWCLSTYVCCRKAYL	0.9548
AMBD9p	MKLLYFLSVAFLVFQAQA	QDEVTAQDEAKAQDELKPKAEDAVMDAENAADNQSPALKPQGSPTDCHRQLGVCRSFLCFFETTIGSCNRHQVCCRRI	0.9355
AMBD10	MKLLYLLGVAFVFTQA	QDGAVAQDEAEADLDEMEEEAEDEFVEAEDAAGMSP ELARKDRPCRKGLFCRPGKQKEHVIGTCPKGLICRIL	0.9352
AMBD11	MKIIYLLGVAFVLSQAQA	QDVVVAQDEAEADLDDIDEEAQDNAMEAEYATMGSPDVKPEYVPCRVLVGVCRPFRCRLNERTIGSCSSNHACCKRY	0.9979
AMBD12	MKFLYLLFGVAFVLTQA	QDIQAQDKAEIQELNQLQLEAKVCLVVMETEHAAADMKYLDPAQPRRRKFCRQGVCKPRCSGNENSSRRCRNHQRCCVRRQ	0.961
AMBD13	MRVLLLLFALLFLVQVQA	QHKAQEEAQDPALQDEAEAVMAAPENTPISRNCNRSGATCRVGFVGFGEIKLGSFAFLRPPCKELPGL	0.9608
AMBD14	MKTPCLLFAVLLVHVQA	MPNPVGEKQPHKEADTWDGVEDDASKAKGNVEAEGAGGENNPMVCSYSGGSCRQRCIGHEVMVGKCYGTFICCVHM	0.8999
AMBD15	MRTLYLLFAVSLFMVQIAPG	FFQIYWNTKCKLNGGSCFLRSCPRQFVSFGTCTQECMCCIR	0.405
AMBD16	MRVLYLLFAVSLMSQLAAG	FPQIGYFHCQQNKQCFQHCIPPNKYIGSCKQLGNCCQRYVQESMGMCCRARRDWFSEV	0.8553
AMBD17	MRILYLLALLFLCQALA	DTLTCTKNNGTCSFMLCPIFMKAIGSCYDGAACKCRRCI	0.8308
AMBD18	MKILYLLVGLFLFLQAASG	LGRCNLLNGVCRHTLCHSLEKYIGRCHRGLRNCCVDDYVLKYKM	0.6229

*Signal peptide cleavage sites on the Beta-defensin showing signal sequence and mature peptide sequence. The probability shows the likelihood that the cleavage site is in that position.*

A4.4 *Crocodylus porosus* exons positions

GENE	Exon 1		Length (bp)	Exon 2		Length (bp)	Exon 3		Length (bp)	Orientation	Total Length (bp)	No of AA
	FROM	END		FROM	END		FROM	END				
CPBD1	46711	46771	61	44024	44145	122				-	183	61
CPBD2	57138	57180	43	56114	56280	167				-	210	70
CPBD3	60316	60373	58	59301	59440	140				-	198	66
CPBD4	89848	89905	58	92446	92564	119				+	177	59
CPBD5	108801	108858	58	110660	110781	122				+	180	60
CPBD6	132653	132710	58	131110	131234	125				-	183	61
CPBD7	144540	144711	172	142603	142772	170				-	342	114
CPBD8	155384	155555	172	153799	153929	131				-	303	101
CPBD9	166677	166806	130	168541	168674	134				+	264	88
CPBD10	190853	190910	58	192815	192948	134				+	192	64
CPBD11	217252	217309	58	218834	218961	128				+	186	62
CPBD12	227551	227608	58	226673	226793	121	225527	225542	16	-	195	65
CPBD13	247061	247118	58	245499	245614	116				-	174	58
CPBD14	254999	255056	58	258285	258418	134				+	192	64

*Positions coordinates of exons. Last codon of CTSB is a marker for the start of the beta-defensin cluster. Orientation of the gene and length of exons and full length of coding sequence.*

A4.5 *Crocodylus porosus* physical properties of mature peptide

GENE	pI	Net Charge	Mr
CPBD1	8.94	4	4790
CPBD2	8.9	4	5469
CPBD3	5.13	-2	5120
CPBD4	7.85	1	3987
CPBD5	5.8	0	4057
CPBD6	9.98	7	4555
CPBD7	5.11	-5	10792
CPBD8	5.64	-4	9318
CPBD9	5.74	-1	7396
CPBD10	6.08	0	4553
CPBD11	9.24	6	4819
CPBD12	9.27	5	5074
CPBD13	8.58	3	4239
CPBD14	9.1	5	5141

*Physical properties of the Crocodylus porosus mature peptide beta-defensins. All properties were achieved by using the protparam program on the ExPASy Server (Gasteiger, E. et al. 2005)*

#### A 4.6 *Crocodylus porosus* signal peptide prediction

	SIGNAL PEPTIDE	MATURE PEPTIDE	Probability
CPBD1	MRLLYLLFAAVMLLFLQAVP	ANGSYYSTLQCRNNHGHCRRLCFHGEQWIGNCNGRHQHCCK	0.8533
CPBD2	MLWFAAFLLAVPGNAQG	SKHVCRTAGGQCRMIGICLSGEVRIGDCFIPVILCCKKYPVRKETGELQGGGA	0.5604
CPBD3	MKFFHLLALLFGIFLATTANG	QRATRYVNHCLQKGGTCRYDDCEAGEEQIGTCYRQTMVCCRDEE	0.9204
CPBD4	MKSLYLILALALFFSQVAPGGA	APLPHEICRSHGGICVANLSLCPHLILQVFGCICCRI	0.7301
CPBD5	MKSLYLILALALFFSQVVPNGG	LPILSFLQCLNLQGTCLLTVGFNGITIRLLGDCDCTP	0.6078
CPBD6	MRILYLLFAVLLFVLQAAPGHG	QPSRSCLDRGGRCIRYNTCHPNLIINARCPHQTVCCRRI	0.8563
CPBD7	MKLLFLLGVTTLVFQAQA	QDVVVAQDKAEPQDLDEMEEAETEVEAQAAGMDFPGLNLGESPACRWRGICRPTHCKKNDPNCRYNPCRFRQERIVGWCLSSHVCCVKAKL	0.9546
CPBD8	MKLLYLLLSVAFVLFQTQA	QDEVLTQDEAKAQDLDELKPKAEDAVMEAVNAADSQSPDLKPHGSPTDCHRLKIGICRHVFCNLFEITIGYCNRRHHVCCRRI	0.9601
CPBD9	MRLLLLSALLFLVLQVQA	QHKAQEEAQDPALQDEAEAVMAAPENTPISRSSRRSGATCRVGFCEGELRLGSCAFLRPCCKELPGL	0.9609
CPBD10	MRFLYLLAVLFFLQVSSG	FVDVAPADTVACRNQGNFCRLGTCPTFEGTGTCNNGALLCCSK	0.4961
CPBD11	MRTLYLLFAVSLFMVQIAPG	FFQIYGNTKLCKLNGGSCFLRSCPFRKFSFGTCTRECMCCIR	0.3998
CPBD12	MRVLYLLFTVSILMLQLAAG	FPKIGYFHCRSQNGNCYQYACPPNTKYIGSCNKLGNCCQRILGGR	0.8976
CPBD13	MRILYLLALLFLCQALA	DTLTCTKNNGTCAFMLCPIFMKAIGTCYDGAAKCCRRI	0.8353
CPBD14	MKILYLLVGLFLQAAASG	LGRCNLLNGVCRHTLCHSLEKYVGRCHRGLRNCCVDDYVLKYKM	0.6228

*Signal peptide cleavage sites on the Beta-defensin showing signal sequence and mature peptide sequence. The probability shows the likelihood that the cleavage site is in that position.*