Whole genome sequencing, de novo assembly, and comparative analysis of *Varanus salvator* and *Megaustenia siamensis*

Miss Wanna Chetruengchai

A Dissertation Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in Biomedical Sciences

Inter-Department of Biomedical Sciences

GRADUATE SCHOOL

Chulalongkorn University

Academic Year 2022

การถอดลำดับสารพันธุกรรมทั้งจีโนม การประกอบขึ้นใหม่ และการวิเคราะห์เชิงเปรียบเทียบของตัวเงินตัวทองและหอยทากบก

น.ส.วรรณนา เชฎฐ์เรืองชัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาชีวเวชศาสตร์ (สหสาขาวิชา) สหสาขาวิชาชีวเวชศาสตร์
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

Thesis Title	Whole genome sequencing, de novo assembly, and comparative analysis of*Varanus salvator* and *Megaustenia siamensis*
By	Miss Wanna Chetruengchai
Field of Study	Biomedical Sciences
Thesis Advisor	Professor VORASUK SHOTELERSUK, M.D.

Accepted by the GRADUATE SCHOOL, Chulalongkorn University in Partial Fulfillment of the Requirement for the Doctor of Philosophy

..... Dean of the GRADUATE SCHOOL

(Associate Professor YOOTTHANA CHUPPUNNARAT, Ph.D.)

DISSERTATION COMMITTEE

..... Chairman

(Wanapinun Nawae, Ph.D.)

..... Thesis Advisor

(Professor VORASUK SHOTELERSUK, M.D.)

..... Examiner

(Associate Professor THANTRIRA PORNTAVEETUS, D.D.S. Ph.D)

..... Examiner

(Associate Professor Monnat Pongpanich, Ph.D.)

..... Examiner

(CHUREERAT PHOKAEW, Ph.D.)

วรรณนา เชฎฐ์เรืองชัย : การถอดลำดับสารพันธุกรรมทั้งจีโนม การประกอบขึ้นใหม่ และการวิเคราะห์เชิงเปรียบเทียบของตัวเงินตัวทองและหอยทากบก. (Whole genome sequencing, de novo assembly, and comparative analysis of*Varanus salvator* and *Megaustenia siamensis*) อ.ที่ปรึกษาหลัก : ศ. นพ.วรศักดิ์ โชติเลอศักดิ์

การเข้าใจโครงสร้างและหน้าที่ของจีโนมสัตว์เป็นสิ่งสำคัญสำหรับการต่อยอดความรู้ในปัจจุบันทั้งทางด้านวิวัฒนาการ ความหลากหลายทางชีวภาพ และการประยุกต์ใช้ด้านเวชสำอางและทางการแพทย์ แต่ละสายพันธุ์มีลักษณะเฉพาะที่เป็นเอกลักษณ์ *Varanus Salvator* เป็นสัตว์เลื้อยคลานชนิดหนึ่งอยู่ในกลุ่มสัตว์มีกระดูกสันหลัง ซึ่งพบแพร่หลายในประเทศไทย และเป็นจิ้งจกที่ใหญ่เป็นอันดับสองของโลก มีวิวัฒนาการมาเป็นเวลาหลายล้านปีเพื่อปรับตัวและอยู่รอดในสภาพแวดล้อมในเมือง คลองที่มีมลพิษ และขยะมูลฝอย นอกจากนี้ยังเป็นสัตว์ที่มีคุณค่าทางเศรษฐกิจโดยเฉพาะหนังซึ่งถูกนำไปใช้ในการผลิตเครื่องหนัง *Megaustenia siamensis* เป็นกลุ่มสัตว์ไม่มีกระดูกสันหลังที่น่าสนใจ มีชื่อเสียงในด้านความสามารถในการผลิตสารยึดเกาะทางชีวภาพ ทำให้สามารถเกาะติดกับใบของต้นไม้ได้แม้ในฝนตกหนัก มีความสำคัญในการประยุกต์ใช้ทางเศรษฐกิจ ชีวภาพเครื่องสำอางและทางการแพทย์ อย่างไรก็ตามข้อมูลจีโนมหรือทรานสคริปโตมของสัตว์ทั้งสองชนิดยังมีการศึกษาน้อย ในการศึกษาครั้งนี้เราจึงประกอบจีโนม(assembly) ที่มีลำดำเบสปริมาณมาก (High-throughput) เพื่อให้ทราบลำดำเบสทั้งหมดของจีโนมโดยไม่มีข้อมูลลำดับเบสอ้างอิงก่อนหน้า (de novo) ผลที่ได้สามารถเป็นข้อมูลอ้างอิงจีโนมของ *V. salvator* และ *M. siamensis* เพื่อเปิดเผยลักษณะทางพันธุกรรม เข้าใจวิวัฒนาการและการปรับตัวของแต่ละสปีชีส์ นอกจากนี้การศึกษาระดับทรานสคริปโตมนำไปสู่การค้นพบโปรตีนที่เกี่ยวข้องกับการสังเคราะห์สารออกฤทธิ์ทางชีวภาพซึ่งมีศักยภาพสูงในการใช้งานด้านเวชสำอางและการแพทย์ ดังนั้นเราสร้างจีโนมโดยการใช้ข้อมูล short-read และ long-read มาวิเคราะห์ร่วมกัน (Hybrid) เพื่อให้ได้จีโนมที่มีคุณค่า นำไปสู่ความเข้าใจทางด้านวิวัฒนาการและการปรับตัว รวมไปถึงยีนที่เกี่ยวข้องกับลักษณะเฉพาะของ *V.salvator* และ *M.siamensis*. นอกจากนี้เรายังศึกษาการแสดงออกของยีนที่แตกต่างกันระหว่างเนื้อเยื่อเท้าและเมนเทิลด้วยเทคนิคทรานสคริปโตม ผลที่ได้พบว่า จีโนมของ *V.salvator* มีขนาด 1.7 Gb N50 71 Mb และ 87.5% ความสมบูรณ์ของการประกอบจีโนม จากการวิเคราะห์จีโนมเปรียบเทียบพบว่า *V.salvator* มีวิวัฒนาการใกล้เคียงกับ *V.komodoensis* เรายังพบยีนที่ควบคุมการแข็งตัวของเลือดและยีนภูมิคุ้มกันโดยกำเนิดซึ่งเป็นลักษณะเฉพาะที่พบใน *V.salvator* จีโนมของ *M.siamensis* มีขนาด 2.59 Gb N50 84.3 Mb และ 85.9% ความสมบูรณ์ของการประกอบจีโนม *M.siamensis* มีวิวัฒนาการใกล้เคียงกับ *Arion vulgaris* ซึ่งจัดอยู่ในกลุ่มหอยทากบก เนื้อเยื่อเท้าและเมนเทิลพบโปรตีนที่เป็นส่วนประกอบหลักของกาว 2 กลุ่ม ได้แก่ โปรตีนกลุ่ม lectin (C-lectin, H-lectin, and C1q) และ โปรตีนกลุ่ม matrilin-like proteins (VWA and EGF) นอกจากนี้ค้นพบ 44 เปปไทด์ต้านจุลชีพ ดังนั้นข้อมูลจีโนมที่ประกอบได้คาดว่าจะให้ข้อมูลที่เป็นประโยชน์สำหรับการศึกษาในอนาคตและมีส่วนอย่างมากในการทำความเข้าใจข้อมูลเชิงลึกทางชีววิทยา นอกจากนี้ข้อมูลทรานสคริปยังเผยให้เห็นองค์ประกอบโปรตีนและสารประกอบออกฤทธิ์ทางชีวภาพซึ่งสามารถนำไปประยุกต์ใช้ทางการแพทย์ต่อไปได้

สาขาวิชา	ชีวเวชศาสตร์ (สหสาขาวิชา)	ลายมือชื่อนิสิต .....
ปีการศึกษา	2565	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6383005020 : MAJOR BIOMEDICAL SCIENCES

KEYWORD: De novo assembly Megaustenia siamensis Transcriptome Varanus salvator Whole genome sequencing

Wanna Chetruengchai : Whole genome sequencing, de novo assembly, and comparative analysis of*Varanus salvator* and *Megaustenia siamensis*. Advisor: Prof. VORASUK SHOTELERSUK, M.D.

Background: Understanding the animal genome structure and function is pivotal for expanding our current knowledge of evolution, biodiversity, cosmeceutical and medical applications. Each species has a unique characteristic feature. *Varanus salvator* is a reptile (vertebrate) prevalent in Thailand. It has evolved over millions of years to adapt and survive in urban environments, polluted canals, and garbage dumps. It still survives without noticeable effects. *Megaustenia siamensis*, an invertebrate, is famous for its ability to produce biological adhesive substances enabling it to stick to trees' leaves even in heavy rain. However, there is very little genomics or transcriptomics data for *V. salvator* and *M. siamensis*. Herein, High-throughput whole-genome assembly would serve as a genomic reference of *V. salvator* and *M. siamensis*. The transcriptomics would lead to the discovery of proteins involving in bioactive compound syntheses, which may have high potential in cosmeceutical and medical applications. Method: We generated reference genomes using a hybrid approach of short-read and long-read sequencing platforms. Performing comparative genomics within and between species to understand the evolution and adaptation. We also studied positively selected genes which were associated with *V.salvator*-specific traits and *M.siamensis*-specific traits. Furthermore, we sought for potential bioactive compounds, antimicrobial peptides, and differential expression between foot and mantle from transcriptomic data. Result: The final size of the reference genome assembly of *V. salvator* was 1.7 Gb with an N50 scaffold size of 71Mb featuring 87.5% completeness of the genome assembly. Our comparative genomic analysis revealed that *V. salvator* is closely related to *V. komodoensis*. We also found a positive selection in *V. salvator* genome within the genes controlling blood clotting and innate immunity genes. The final size of the reference genome assembly of *M. siamensis* was 2.59 Gb with an N50 scaffold size of 84.3Mb featuring 85.9% completeness of the genome assembly. The phylogenetic tree showed that *M. siamensis* was most closely to *A. vulgaris*, which is a group of land snails. Positive selection was found in pathways related to ubiquitin-proteasome. Moreover, RNA from foot and mantle tissues suggested the major components of the glue, which comprised lectin-like proteins (C-lectin, H-lectin, and C1q) and matrilin-like proteins (VWA and EGF). The result also reveals 44 potential antimicrobial peptides (AMPs). Conclusion: The draft genome data are expected to provide useful information for future studies and contribute significantly to understanding the biological insight. Furthermore, the transcriptome data reveal the highest protein composition and bioactive compounds, which might be useful for medical applications.

Field of Study:	Biomedical Sciences	Student's Signature .....
Academic Year:	2022	Advisor's Signature .....

ACKNOWLEDGEMENTS

I would like to thank Excellence Center for Genomics and Precision Medicine, King Chulalongkorn Memorial Hospital, The Thai Red Cross Society for giving me the opportunity to be a student here, and the Second Century Fund (C2F), Chulalongkorn University for providing the funding for the study. Very special gratitude goes out to all down at the Second Century Fund (C2F), Chulalongkorn University, for providing the funding for the study, and This research was funded by Health Systems Research Institute (65-040), TSRI Fund (CU_FRB640001_01_30_10).

I would like to express my sincere gratitude to my advisor Prof. Vorasuk Shotelersuk for the continuous support of my Ph.D. study and related research, for his valuable suggestion, his encouragement, his motivation, his confidence, his listening, and his helpfulness in all concerns, which helped us me to the successful accomplished of my research and facilitated the realization of this book. I could not have imagined having a better advisor for my study.

I would also like to thank Associate Professor Kornsorn Srikulnath's team, the Animal Genomics and Bioresources Research Center (AGB Research Center), Faculty of Science, Kasetsart University, Bangkok, for providing a Varanus salvator sample, and Professor Somsak Panha's team, Animal Systematics Research Unit, Department of Biology, Faculty of Science, Chulalongkorn University, for supporting Megaustenia siamensis sample.

I would also like to thank the member of the Excellence Center for Medical Genetics, King Chulalongkorn Memorial Hospital, for listening and assisting, giving me valuable suggestions and encouragement. Last but not least, I thank my family for their help throughout this research. I would not have succeeded in this thesis without all of their help and support.

Wanna Chetruengchai

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Background and Rationale

The advances and developments in high throughput sequencing (HTS) technologies, also known as Next-generation sequencing (NGS), lead to obtaining massive datasets at a low cost and in a short time (1). This revolutionized the studies of genome sequencing (genomics) used in human and animal genetics. The field of genome assembly is thriving, with new genomes of all kinds of organisms becoming more available. Our 11,032 metazoan (multicellular animals) assemblies are available in the GenBank database (accessed on November 4th, 2022). De novo genome assembly is one of the most challenging computational processes in modern genomics because without a reference genome. It is constructed from DNA fragments (short/long DNA sequences) without knowledge of the length, location, or composition for determining the sequence composition of the genetic material (DNA) within the cell of an organism. Understanding the animal genome structure and function is pivotal for expanding our current knowledge of evolution, biodiversity, and medicine. The animal kingdom is divided into two groups: those with a presence backbone (vertebrates) and those without a presence backbones (invertebrates). *Varanus salvator* is a reptile (vertebrate) prevalent in Thailand. They can adapt to urban environments such as city parks, roads, garbage dumps, and polluted canals, and they also eat any animal, carrion, or food waste (2). It still survives without noticeable effects and currently is a valuable economic animal with its leather being used for the production of leatherwork. *Megaustenia siamensis* is an interesting invertebrate group that has completely invaded the land and has several specific adaptations to diverse terrestrial habitats. Subsequently, the transcriptomics of some selected land snails with high potential in further application will become the foundation for further genomic analyses of other land snails in the region. This includes the discovery of proteins involving in bioactive compound syntheses which will have high potential in cosmeceutical, and medical applications. However, the study of the bioactive compound components in mucus from most land snail

species, especially in Thailand, has received less attention. The impact of this research will lead to the further test of the mucus from different species of land snails for its biological and therapeutic properties that will show high efficacy of biological substances for prospective cosmetic and skin care applications. Herein, High-throughput whole-genome assembly can serve as a genomic reference of *V. salvator* and *M. siamensis* to uncover genetic features and understand both evolution and adaptation in the species. The transcriptomics will lead to discovery of proteins involving in bioactive compound syntheses which will have high potential in cosmeceutical, and medical applications

Keyword

De novo assembly, *Megaustenia siamensis,* Transcriptome, Varanus salvator, Whole genome sequencing

Research questions and hypotheses

Thailand has a unique geography and environment. The unique physical and biological features of Thai indigenous animals which allow them to adapt and survive in their ecosystem should come from their DNA sequences. We propose to study two indigenous animals, which are representative of vertebrates and invertebrates. Each animal has specific research questions and hypotheses as follows.

1. Research question of *Varanus salvator*

How different of its genome from other reptiles?

What are the genes responsible for Varanus's unique immune system?

1.1 Hypothesis of *Varanus salvator*

Varanus salvator have unique genes for its immune system.

2. **Research question of *Megaustenia siamensis***

How different of its genome from other snails?

What are the genes encoding for its biological adhesive substance?

2.2 Hypothesis of *Megaustenia siamensis*

M. siamensis have genes involving in biological adhesion.

Objectives

1. To generate high-quality genome sequence of Asian water monitor (*V. salvator*) and Land snail (*Megaustenia siamensis*) through whole genome sequencing and de novo genome assembly.

2. To study evolution through comparative genomics within and between species.

3. To gain biological insight from the two newly assembled genomes.

Research design

In silico and descriptive studies

Ethical consideration

The *V.salvator* study was reviewed and approved by the Sri Nakhon Khuean Khan Park (Royal Forest Department, Ministry of Natural Resources and Environment) and Kasetsart University (0909.6/15779).

The *M.siamensis* study was reviewed and approved by the the Animal Care and Use Committee of Faculty of Science, Chulalongkorn University (Protocol Review No. 2123023)"

Graphical abstract

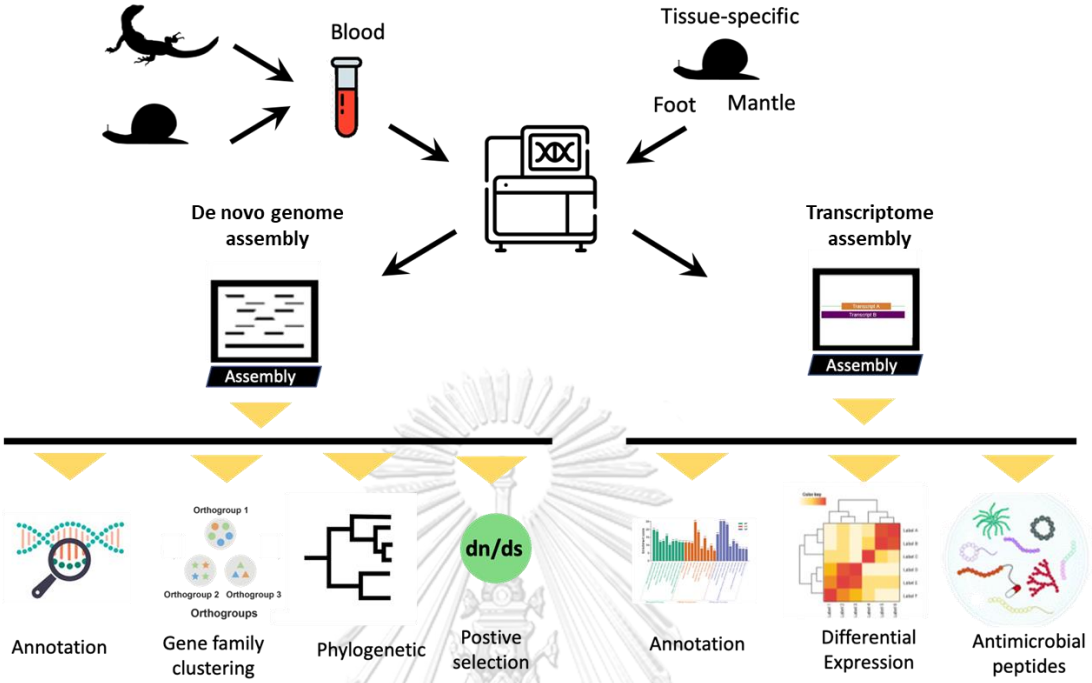

Figure 1 Data analysis overview

CHAPTER 2

REVIEW LITERATURE

2.1 Whole genome sequencing for de novo assembly

The high throughput DNA sequencing technology can still not decipher the entire genetic material of the target organism in one piece. Therefore, the target organism's genome is broken up into millions of DNA fragments (also called reads). The length of reads depends on the sequencing platform, such as short-read or long-read sequencing.

Illumina sequencing platform is predominant in short read sequencer technology. Generally, generating read lengths between 2*150 to 2*300 bp has an errors rate between 0.02%–0.05% (3). However, short-reads are still challenges in de novo assembly. Their limited ability to cover a repetitive or heterozygous sequence was incomplete and heavily fragmented (4). Long reads can help this problem. Pacific Biosciences (PacBio) is long-reads sequencing platform that generates read lengths up to 10 kb. However, Long-reads are sufficient to cover repeat regions, but this technology suffers from high costs and error rates (~15%) (5). The de novo assembly approach can combine data from long and short reads (also called hybrid approach). Short reads correct the sequencing error from long reads to increase the assembly quality. However, the sequencing technology and hybrid approach still incomplete genome (such as a lot of gaps or fragments). The sequencing technology has additional techniques. 10X genomics linked read sequencing use barcode to tag short reads originating from the same long DNA fragment. Hi-C sequencing data is highly recommended for de novo assembly because it helps genome completion. Hi-C scaffolders use Hi-C data to connect contigs from linkage information. These technologies are increasingly used to enhance genomic assemblies (4). Figure 2 show type of sequencing reads from different genomic technologies to construct the genome assembly (6).

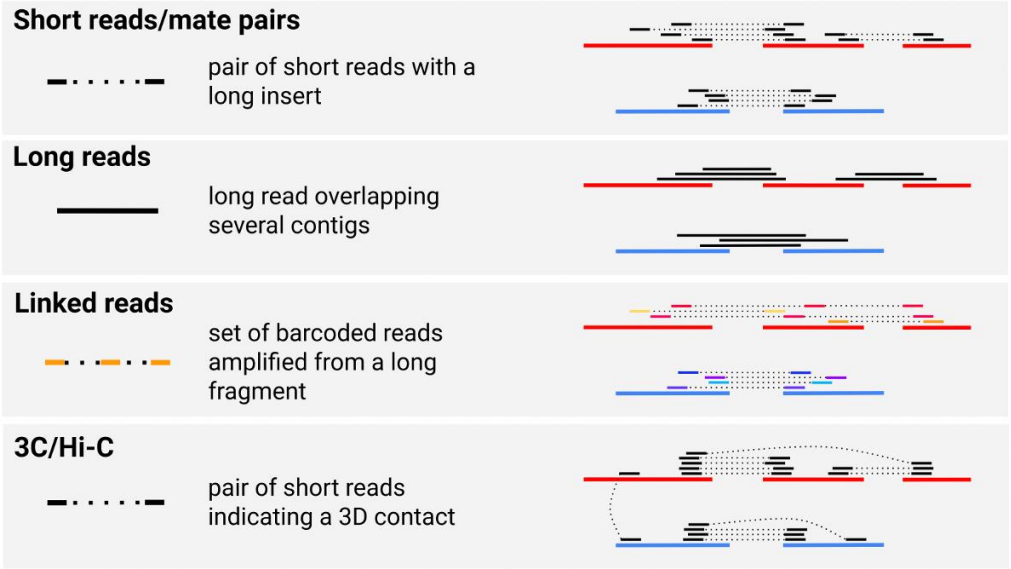

Figure 2 Sequencing technologies in this study

2.2 De novo genome assembly approach

The DNA of the target organism is broken up into millions of DNA fragments (also called reads). The reading length depends on the sequencing platform used, such as short-read or long-read sequencing. The DNA fragmented sequences must be assembled by finding shared read regions to construct a genome-level structure. The basic strategy for de novo assembly; First, reads were joined together based on overlapping regions to form a long consensus sequence without gaps (called contigs). Second, the contigs that can be linked are grouped by using additional information (N letters), also called gaps. Then, the set of connected contigs is defined as a scaffold (4) (Figure 3). However, a variety of gaps in scaffolds result in genomes that are far from complete.

2.2.1 Assembly approaches

In the first step, k-mer frequency is calculated to determine the genome characteristics (genome size, repeat content, sequencing error rates, and rate of heterozygosity) from raw read data using jellyfish software and visualized by Genomescope (7). K-mer frequency distribution is a powerful approach to using raw Illumina DNA shotgun reads and then following the de novo genome assembly workflow: NGS sequences --> Assembly (To

generate contigs from fragmented sequence reads) --> Scaffolding or merging (To increase assembly continuity (e.g., N50 length by decreasing scaffold numbers) --> Correction/Polishing using raw Illumina DNA shotgun reads (To improve assembly accuracy) --> Assessment and decision (To determine assembly quality and completeness) (8).

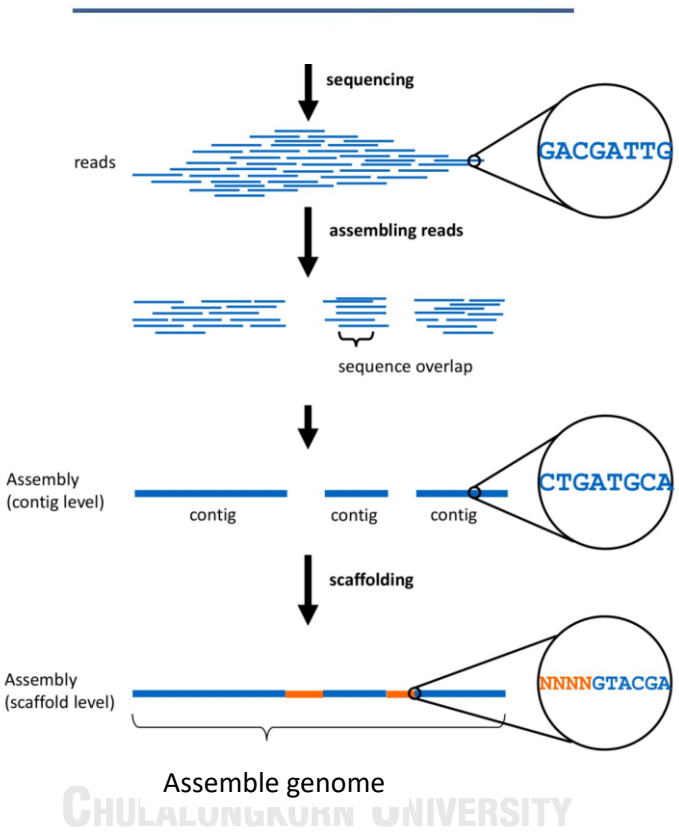

Figure 3 The general steps of de novo assembly

2.2.2 Check the assembly quality before annotation

A critical step in genome assembly is estimating draft assemblies' quality. QUAST (QUality ASsesment Tool for Genome Assembly) is the most popular software to assess the assembly size, number of contig or scaffold, N50, GC content, and number of gaps) (9). N50 is one of the most popular metrics, which is defined as the length of the largest contig (or scaffold) that covers at least half (50%) of the total base content of the assembly. Benchmarking Universal Single-Copy Orthologs (BUSCO) is a software for quantitative assessment of genome

assembly (10), based on the concept of single-copy orthologs. The output provides how many universal genes matches complete with input data, duplicated, fragmented, or missing genes. Many missing or fragmented BUSCO genes in an assembly indicates low quality. The high quality of genome depends on several factors such as sequencing platform, depth of coverage, and software.

2.2.3 Genome annotation

Annotation uses identifying and describing regions of biological interest within a genome, which includes two main levels. First, structural annotation identifies genomic elements (Introns/exons, CDS, stop, start, amino acid sequence). Second, Functional annotation is attaching biological information to genomic elements. Gene ontology (GO) is the most widely used for identifying gene function, providing three categories: biological processes (BP), molecular functions (MF), and cellular components (CC) (11).

2.2.4 Comparative genomics

Comparative genomics help to identify genes that are unique or shared with different species, such as clustering of orthologous groups. It also provides a powerful tool for studying evolution, such as phylogenetic tree inference or positive selection (12).

Orthologue identification is critical in comparative genomics. Orthologues are genes that evolved from a common ancestor gene through speciation. The term orthologue most commonly refers to genes with conserved functions that retain the same function throughout evolution. In addition, orthologue is a key concept in phylogenomics. Most phylogenies require orthologous, rather than paralogous, relationships, particularly one-to-one (1:1) orthologous relationships, which means that species have not undergone gene duplication since their divergence (13). Estimating nonsynonymous/synonymous rate (dN/dS) Specifically, under positive selection (dn/ds >1) is crucial for the evolution of new functions or differences among specie from protein changes (14).

2.3 Transcriptome sequencing (RNA-Seq)

Using high-throughput sequencing techniques (Next-generation sequencing, NGS), transcriptome sequencing (RNA-Seq) is a recently developed technology that identifies the sequence of all RNA transcripts in a given specimen (15). RNA-Seq can characterize and quantify the transcriptome with greater accuracy and detail, and it can also quantify differences in transcript expression between samples. The process of analyzing RNA-Seq data involves several steps, including quality control, data preprocessing, transcriptome assembly (either reference-guided or de novo, depending on the presence of a reference genome), quantification, statistical analysis, and functional annotation (16). De novo transcriptome assembly uses the redundancy of the sequencing reads themselves to reconstruct overlapping RNA-Seq reads into transcripts (17). Understanding which genes and their levels of expression are expressed in various specimens is another goal of differential expression analysis.

2.4 *Varanus salvator*

Water monitors (*Varanus salvator macromaculatus*, Deraniyagala 1944) are the largest monitor in Southeast Asia and the second largest lizard globally after the Komodo dragon (18). They are a member of the family Varanidae within the genus *Varanus*. They live in semi-aquatic ecosystems inhabiting wetlands, hills, mangroves, and canals. As a prolific scavenger, the water monitor consumes animal, carrion, or food waste. This results in many beneficial effects including removing many infection sources from the environment, helping to clean surroundings, maintaining the ecosystem, and balancing the food chain. The gut microbiota of *V. salvator* reveals the enrichment of bactericidal activity components (19). Although *V. salvator* lives in a contaminated environment and consumes decaying carcasses, it still survives without any noticeable effects, suggesting its strong immunity. While mammals rely more on adaptive or acquired immunity, these reptiles rely heavily on innate immunities to combat infections (20). Komodo dragon have a positive advantage of a blood-clotting mechanism, which is a part of their innate immune system helping them to survive in extreme environments (21,

22). The transposable elements (TEs) within the vertebrate genome shaped innate immune evolution by triggering the antiviral innate immune pathway (23). However, the immunity of *V. salvator* is still unknown.

2.5 *Megaustenia siamensis*

Thailand is a biodiversity hotspot with a diverse range of snails (1). Land snails are one of the interesting groups that adapted to various terrestrial habitats and are increasingly used in medical applications. Snail slime (mucus) performs a variety of functions in the animal, including adhesive, emollient, moisturizing, lubricant, and defense, and it includes several bioactive components, including antibacterial and antioxidant compounds (2). For example, the consequential result of some analyses on the mantle mucus secretion from *Hemiplecta disctincta* for skin care processes and application to cosmetic industry lead to the new natural cosmetic products which were officially reported in The Wall Street Journal in February-March 2015 issue, and in a Japanese magazine. However, studying the genomic structure and bioactive compounds of land snails in Thailand has received less attention. With the development of high-throughput sequencing technologies, we combine genome and transcriptome data to uncover genetic features, understand both evolution and adaptation, and discover potential molecular genetics, which is critical for increasing our existing knowledge of evolution, biodiversity, and medicine.

CHAPTER 3

MATERIALS AND METHODS

3.1 DNA sequencing

3.1.1 DNA sequencing of *V. salvator*

High-quality genomic DNA was extracted from *V. salvator*, A total of 135 Gb was sequenced using the Single Molecule, Real-Time (SMRT) Sequencing, according to each manufacturer's instruction. A total of 403 GB (2 x 150 bp) was generated by 10x Genomics barcoded library and sequenced on the Illumina NovaSeq6000.

3.1.2 DNA sequencing of *M. siamensis*

Libraries for PacBio's SMRT (single molecule real time) sequencing and 10x Genomics Chromium linked-reads were constructed as per the manufacturers' instructions using high-molecular weight (HMW) DNA extracted from *M.siamensis*. A total of ~200 GB Pacbio and Illumina (2 × 150 bp) data were generated. Additionally, 200 GB was generated by Hi-C library and sequenced on the Illumina HiSeqX.

3.2 Genome size estimation

The genome size was estimated by K-mer frequency-based Jellyfish of the raw illumina sequence reads, and the histogram was uploaded to GenomeScope to estimate genome size, repeat content, and heterozygosity (7).

3.3 De novo Genome Assembly

3.3.1 De novo genome assembly of *V.salvator*

A total of 403 GB with 10X genomics linked read were assembled with Supernova software (version 2.1) by default parameters and available GitHub: https://github.com/10XGenomics/supernova-chili-pepper, which reads were demultiplexed, converted and build a graph-based assembly and then using pseudohap's parameter to generate scaffold in fasta format.

A total of 150 GB with PacBio, DEXTRACTOR (https://github.com/thegenemyers/ DEXTRACTOR) was used to correct the Pacbio long-reads bam file (pulls sequence, Quiver, and/or Arrow information) and convert to FASTA format files. Sequences shorter than 1000bp and quality (QV) < 0.8 were filtered out. The clean reads were used for de novo assembly with CANU pipeline to generate the initial draft assembly comprised of 3 steps: correction, trimming, and assembly (using default parameters) (24).

Quickmerge integrated the supernova scaffolds and contigs from Pacbio with the default settings (25). Redundancy reduction, scaffolding, and gap closing were carried using the Redundans pipeline (25). The draft assembly was polished by aligning Illumina reads to the genome using bwa, using the "mem" option (26). The assembly was polished with Pilon (27), resulting in the final assembly. The assembly metrics visualized by circular plot (https://github.com/rjchallis/assembly-stats).

3.3.2 De novo genome assembly of *M.siamensis*

Approximately 200 GB with PacBio, DEXTRACTOR (https://github.com/thegenemyers/ DEXTRACTOR) was used to correct the Pacbio long-reads bam file (pulls sequence, Quiver, and/or Arrow information) and convert to FASTA format files. Sequences shorter than 1000 bp and quality (QV) < 0.8 were filtered out. The clean reads were used for de novo assembly with CANU pipeline to generate the initial draft assembly comprised of 3 steps: correction, trimming, and assembly (using default parameters) (24).

Subsequently, short reads were used to polish by Pilon and removed redundant sequences in the assembly by purge_dups. Hi-C reads were mapped to de novo assembled contigs to construct contacts among the contigs using bwa. Then, the final assembly used HiRise scaffolding method to link contigs together. The assembly metrics visualized by circular plot (https://github.com/rjchallis/assembly-stats).

3.4 Genome assembly evaluation

The genome assemblies were quality assessed using descriptive measures, e.g., numbers of contigs, total number assembled bases, and completeness, implementing analyses tools QUAST. The Benchmarking Universal Single-Copy Orthologs pipeline (BUSCO) was used to evaluate the completeness of the genome to obtain the percentage of single-copy orthologous with vertebrata version 10, including 3,354 genes for *V. salvator* and metazoan version 10, including 954 genes for *M.siamensis.*

3.5 Genome annotation

De novo repeat annotation was created by RepeatModeler (http://www.repeatmasker.org/RepeatModeler/). We then used the sequence libraries as a query to mask repetitive elements with RepeatMasker (http://www.repeatmasker.org). We used the "calcDivergenceFromAlign.pl" script in the RepeatMasker pipeline to calculate the Kimura divergence values and plotted the repeat landscape with repeats presented in genomes. The protein-coding genes were annotated using the MAKER pipeline

Gene function annotation was performed using InterProScan to search for domains or motifs in public databases and the web-based platform Kyoto Encyclopedia of Genes and Genomes (KEGG). Orthology assignments and predictions of KEGG pathways were carried out through the KEGG Automatic Annotation Server (KAAS) using bi-directional best hit (BBH) BLAST method (https://www.genome.jp/kegg/kaas/). Similarly, the draft genomes were also scanned for clusters of orthologous groups (COGs) annotations using eggNOG-mapper v2 (http://eggnog-mapper.embl.de).

3.6 Phylogenetic and comparative analysis

3.6.1 Phylogenetic and comparative analysis of *V. salvator*

Phylogenetic tree constructed using one-to-one orthologous genes , which were analyzed with OrthoFinder (28) by comparing with 19 reptile species (*Alligator mississippiensis*, *Alligator sinensis*, *Anolis*

carolinensis, *Chelonia mydas*, *Chrysemys picta*, *Crocodylus porosus*, *Deinagkistrodon acutus*, *Dopasia gracilis*, *Eublepharis macularius*, *Gavialis gangeticus*, *Gekko japonicas*, *Ophiophagus hannah*, *Pelodiscus sinensis*, *Pogona vitticeps*, *Protobothrops mucrosquamatus*, *Python bivittatus*, *Shinisaurus crocodilurus*, *Varanus komodoensis*, and *Varanus salvator*), three avian species (*Taeniopygia guttata*, *Meleagris gallopavo*, and *Gallus gallus*) and four mammals (*Mus musculus*, *Canis familiaris*, *Homo sapiens*, *and Ornithorhynchus anatinus*). The one-to-one orthologous proteins were aligned with PRANK (29). The phylogenetic tree construction using RAxML v8.2.8 with 1000 bootstrap and the phylogenetic tree was visualized by IQ-TREE (30). The evolution time obtained from Timetree database (http://www.timetree.org/).

Additionally, a subset genome in a clade of *V. salvator* (including 5 genomes; *V. komodoensis*, *S. crocodilurus*, and *D. gracilis*, *A. carolinensis* and *P. vitticeps*) were used to estimate species-specific and presence/absence of gene families among the six genomes using Orthovenn2 (31).

3.6.2 Phylogenetic and comparative analysis of *M. siamensis*

Phylogenetic tree constructed using one-to-one orthologous genes, which were analyzed with OrthoFinder by comparing with 10 gastropod species (*Achatina fulica*, *Aplysia californica*, *Arion vulgaris*, *Biomphalaria glabrata*, *Candidula unifasciata*, *Chrysomallon squamiferum* (Scaly-foot gastropod), *Gigantopelta aegis*, *Haliotis rufescens* (red abolone), *Lottia gigantea*, and *Pomacea canaliculate*), and 3 bivalvia species (*Crassostrea gigas* (Pacific oyster), *Dreissena polymorpha*, and *Mizuhopecten yessoensis*). The one-to-one orthologous proteins were aligned with PRANK. The phylogenetic tree construction using RAxML v8.2.8 with 1000 bootstrap. Divergence time was computed using the MCMCTREE program implemented in the PAML v4.8 package with the correlated molecular clock. We applied four fossil calibration point for estimation of divergence times. The fossil of Gastropoda and Bivalvia divergence time (515.0-541.7 MYA), Caenogastropoda and Heterobranchia (238.6-429.9 MYA), *Pomaceae canaluculata* and *Lottia gigantea* (238.6-429.9 MYA), and *Octopus bimaculoides* (480.0-559.4 MYA) for the root. The calibration point obtained from Timetree database

(http://www.timetree.org/)(17). The phylogenetic tree was visualized by FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

M.siamensis, *A.vulgaris*, *C.unifasciata*, and *A.fulica* were used to estimate species-specific and presence/absence of gene families by Orthovenn2.

3.7 Positive selection analysis

The function of positive selection is not well understood, particularly the evolution of host adaptation and specialization. One-to-one orthologues proteins were aligned using PRANK, and codon alignments were generated using PAL2NAL (32). The aBSREL model (adaptive Branch-Site Random Effects Likelihood) implemented in the Hyphy package (v2.5.15) was used to test if positive selection had occurred on a proportion of branches (dn/ds >1) with P-value < 0.05. We further employed Human NCBI EntrezGene IDs to reference human genes with signatures of positive selection. GO enrichment and functional groupings were determined by Fisher's exact tests ($p < 0.05$) from PANTHER databases (33).

4. RNA sequencing (Transcriptome sequencing)

RNA was extracted from foot and mantle tissue. Libraries were produced using TruSeq Stranded mRNA, and sequenced on by Illumina Novaseq with 150 bp paired-end. Approximately 30 GB of foot and mantle tissues had been assembled using both de novo and genome-guided approaches with Trinity (34). Redundant transcripts were identified and removed from the assembly via CD-HIT-EST (35)with an identify threshold of 98% sequence similarity. RNA-Seq by the Expectation-Maximization (RSEM) was used to analyze transcript quantification (36). Normalization and differential expression analysis were carried out using EdgeR (37) with FDR correction ≤ 0.05, P value ≤ 0.01, fold change ≤ 2 or ≥ 2. TransDecoder was used to identify candidate coding regions within transcript sequences (https://github.com/TransDecoder/TransDecoder/wiki). Finally,

Trinotate was used for functional annotation of the transcriptome by performing BLASTX searches on the Swiss-Prot databases to generate gene ontology terms (http://trinotate.github.io).

The signal peptide was also identified by SignalP 5.0 (38). Candidate peptides identified as antimicrobial show all four statistical approaches as positive by mapping with the Collection of Anti-Microbial Peptides (CAMP) using four machine-learning algorithms: Support Vector Machine (SVM), Discriminant Analysis (DA), Artificial Neural Network (ANN), and Random Forest (RF) (39). Moreover, the iACP online tool (https://lin.uestc.edu.cn/server/iACP) was used to predict the anticancer activity of the identified peptides.

CHAPTER 4

RESULTS & DISCUSSION

4.1 Varanus salvator

4.1.1 De novo genome assembly of *Varanus salvator*

The genome size of *Varanus salvator* is estimated to be around 1.5 Gb with 0.185% heterozygosity by raw data of linked-read sequencing data (Figure 4A). We obtained 135 Gb of raw PacBio long reads. Sequences shorter than 1000bp and quality (QV) < 0.8 were filtered out. Remaining 111 Gb were used for de novo assembly with CANU software. It resulted in an assembly of ~1.89 Gb containing a total of 4,006 contigs with an N50 contigs length of 8.8 Mb. All 403 GB raw data of linked-read sequencing data were then assembled using Supernova Assembler. The results showed 1,443 scaffolds with an N50 contigs length of 43 Mb and a total sequence length of 1.57 Gb. Additionally, the supernova scaffolds and contigs from Pacbio were integrated with the default settings by Quickmerge. Redundancy reduction, scaffolding, and gap closing were then carried out using Redundans pipeline. The draft assembly was polished by aligning 10x Genomics Illumina reads to the genome using bwa-mem v0.7.5 and Pilon v1.22 resulting in the final assembly. The final draft genome assembly contained 858 scaffolds with an N50 scaffold length of 71 Mb (longest scaffold: 121 Mb). The GC content of the water monitor is 44.0%. Detailed genome statistics are presented in Figure 4B. The *V. salvator* genome assembly is 1.70 Gb in size, which is ~12% bigger than the genome of the Komodo dragon (*Varanus komododensis*) from the same genus and ~5% smaller than the green anole (*Anolis carolinensis*), a squamate lizard model (Supplementary Table S1).

The Benchmarking Universal Single-Copy Orthologs pipeline [BUSCO v4.0.2] was used to evaluate the completeness of the genome and obtain the percentage of single-copy orthologous with vertebrata_odb10 BUSCO set (3,354 gene vertebrata gene set). 2,933 (87.5%) of the 3,354 complete expected vertebrata genes were identified as complete, including 2,793 (83.3%) single-copy and 140 (4.2%) duplicated. 135 (4.0%) fragmented

vertebrates were present, possibly due to incomplete assembly, and only 286 (8.5%) genes were considered missing in the genome assembly (Table 1).

	Varanus salvator
Assembly size (bp)	1,702,614,977
Number of scaffolds	858
Scaffold N50	71,461,993
Number of protein-coding genes	21,937
Repeat content, %	37.43
GC content (%)	44.03
Complete BUSCO, %	87.5
Complete and Single-copy BUSCO (%)	83.3
Complete and Duplicated BUSCO (%)	4.2
Fragmented BUSCO (%)	4.0
Missing BUSCO (%)	8.5
Total number of metazoa_odb10	3,354

Table 1 Genome statistics of *V.salvator*

The completeness of the *V. salvator* assembly was comparable to other published reptile genome assemblies (Supplementary Table S1). BUSCO was also run with the same parameters on 18 reptile genomes for comparative analyses (*Alligator mississippiensis, Alligator sinensis, Anolis carolinensis, Chelonia mydas, Chrysemys picta, Crocodylus porosus, Deinagkistrodon acutus, Dopasia gracilis, Eublepharis macularius, Gavialis gangeticus, Gekko japonicas, Ophiophagus hannah, Pelodiscus sinensis, Pogona vitticeps, Protobothrops mucrosquamatus, Python bivittatus, Shinisaurus crocodilurus and Varanus komodoensis*). The result showed that our *V. salvator* draft genome has good quality and a high completeness level. Other previously reported standard genomes are presented (Figure 4C).

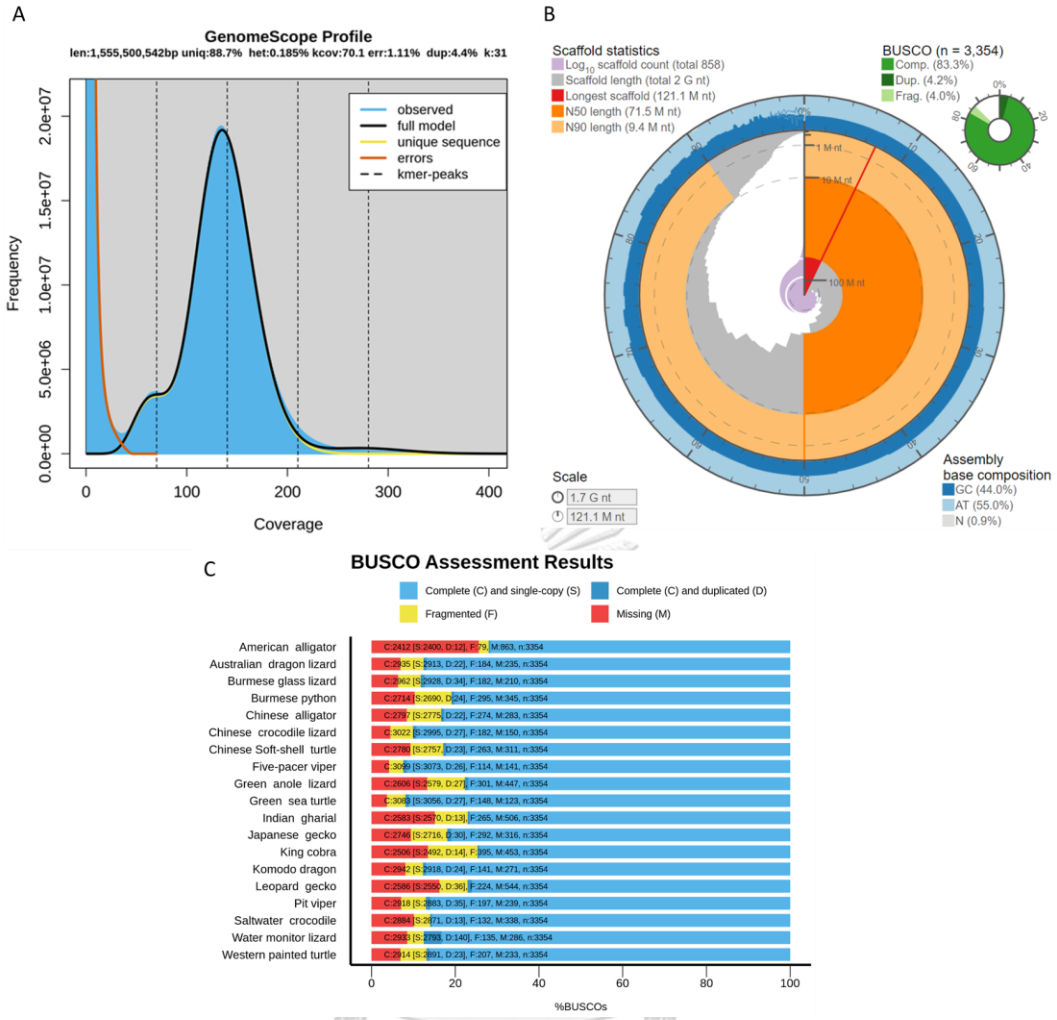

Figure 4 Genome estimation (A), Genome assembly (B), and comparison genome completeness (C) of *V.salvator*

4.1.2 Genome annotation

We utilized the RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) for run de novo repeat annotation and constructed species-specific repeat sequence libraries for the *V. salvator*. We then used the sequence libraries as a query to mask repetitive elements with RepeatMasker (http://www.repeatmasker.org). We found that repetitive elements accounted for 37% of the *V. salvator* genome (Figure 5A and Supplementary Table S2). The proportion of LINEs in the *V. salvator* genome is 16.76% (the total TE is 35.99%) (Supplementary Table S6) which is higher than other squamate relatives such as *V. komodoensis* (LINE; 13.43% of the 32.18%) and *A. carolinensis* (LINE; 12.19% of the 33.82%).

For protein-coding gene analysis, we identified 21,937 protein-coding genes based on the combination of ab initio gene prediction by MAKER version 2.31.10 pipeline and homology search by SNAP and Augustus version 3.3.1. We were able to get functional annotations from 14,668 of 21,937 protein-coding genes. COG annotated could be group into 25 functional categories (Figure 5B). The largest category was Cellular process and signaling (8,171 COG annotations, 37.25%), containing [T] Signal transduction mechanisms is the predominant functional category (19.18%). The other COG functional categories consisted Information storage and processing of 15.26% ([K] Transcription (7.78%)) and Metabolism of about 14.35% ([P] Inorganic ion transport and metabolism) of the COG categories. In addition, 5,910 (26.94%) genes belonged to the "Function unknown" category and 1,359 (6.20%) not related in COG (Figure 5B). According to KEGG pathway analysis (Figure 5C). Signal transduction (1783 genes) pathway constitutes the predominant pathways of Environmental Information Processing including 413 MAPK signaling pathway owning the most annotated.

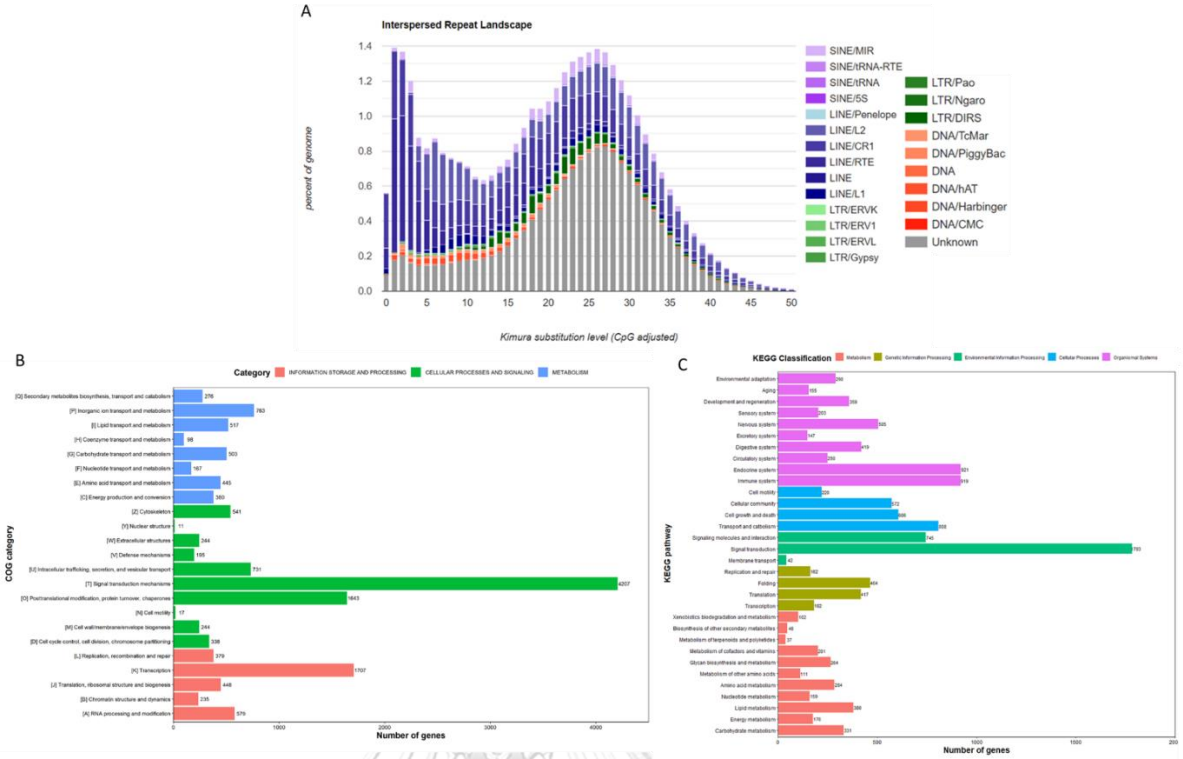

Figure 5 Interspersed Repeat Landscape (A). Functional classification in Clusters of Orthologous Groups of proteins (COG) (B) and KEGG (C) in the *V.salvator* genome

4.1.3 Phylogenetic and comparative analysis

We constructed a phylogenetic tree to understand the genome evolution of *V. salvator* by IQ-TREE with 1,000 bootstrap. The IQ-TREE input data was the multiple sequence alignment from the PRANK using 1,339 single-copy orthologous genes, which were analyzed with OrthoFinder version 2.3.12 by comparing the whole genome sequences of 19 non-avian reptile species, four mammals (*M. musculus*, *C. familiaris*, *H. sapiens*, and *O. anatinus*), and three avian species (*T. guttata*, *M. gallopavo*, and *G. gallus*). According to the tree, *V. salvator* was most closely related to *V. komodoensis* with high bootstrap support (100%); both are the *Varanus* genus of the Varanidae family (Figure 6A). The time of divergence between *V. salvator* and *V.komodoensis* was estimated to be around 55 MYA, obtained from TimeTree database (Figure 6B).

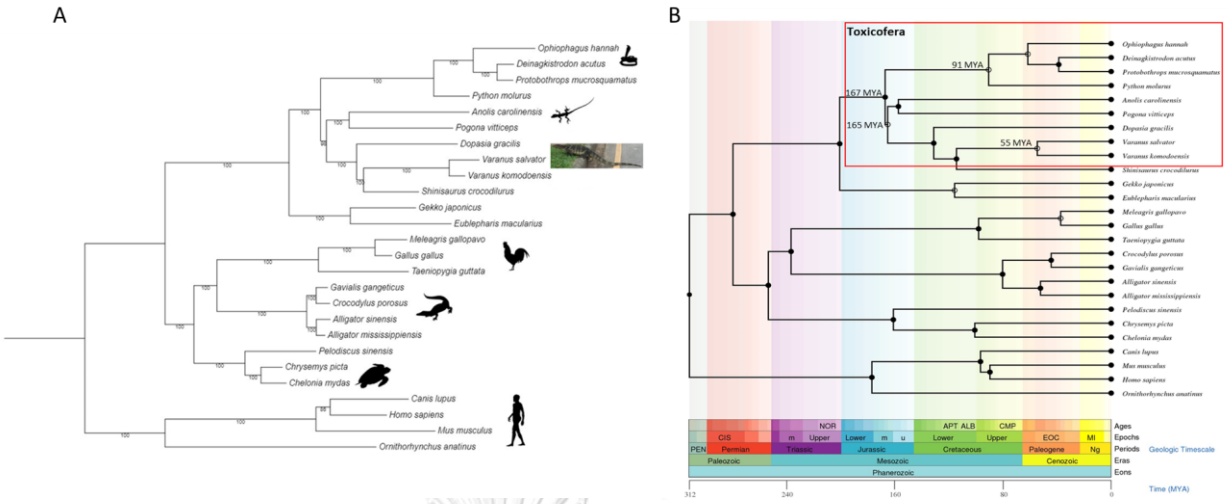

Figure 6 one to one orthologous phylogenetic tree (A). Evolutionary Timetree (B).

4.1.4 Gene family clustering

we use Orthovenn2 to identify species-specific genes, presence/absence of gene families, and pathway enrichment among the six genomes of the same clade as *V.salvator* including four genomes of the Anguimorpha suborder (*V. salvator* ,*V. komodoensis*, *S.crocodilurus*, and *D. gracilis*), and two members of the squamata order (*A. carolinensis* and *P.vitticeps*). The selected genome formed 21,221 clusters of 18,687 orthologous clusters (containing two species at least) and 2,534 single-copy gene clusters. The Venn diagram shows that 7,861 gene families were shared among the six genomes supporting their conservation in the lineage (Figure 7A). A total of 234 out of 7,861gene families are specific to *V. salvator*. We were able to annotate 94 out of 234 clusters according to the Swiss Protein Database, including 77 clusters (Supplementary Table S3) in biological processes (BP), 16 in molecular functions (MF), and 1 in cellular components (CC). For GO enrichment analysis, the gene family encoding vomeronasal type 2 receptors (V2Rs) played a significant role in response to pheromones (GO:0019236, GO level: BP, P = 1.37e-11). The histocompatibility antigen is essential for immune response (GO:0006955, GO level: BP, P = 3.89e-09) (Supplementary Table S4). Figure 7B shows the pairwise heatmap based on the similarity matrix, which illustrates the overlapping cluster numbers for the six genomes. Both *V.*

salvator and *V. komodoensis* shared the highest number of orthologue clusters (18,199). All results emphasized the close relationship between *V. salvator* and *V. komodoensis*.

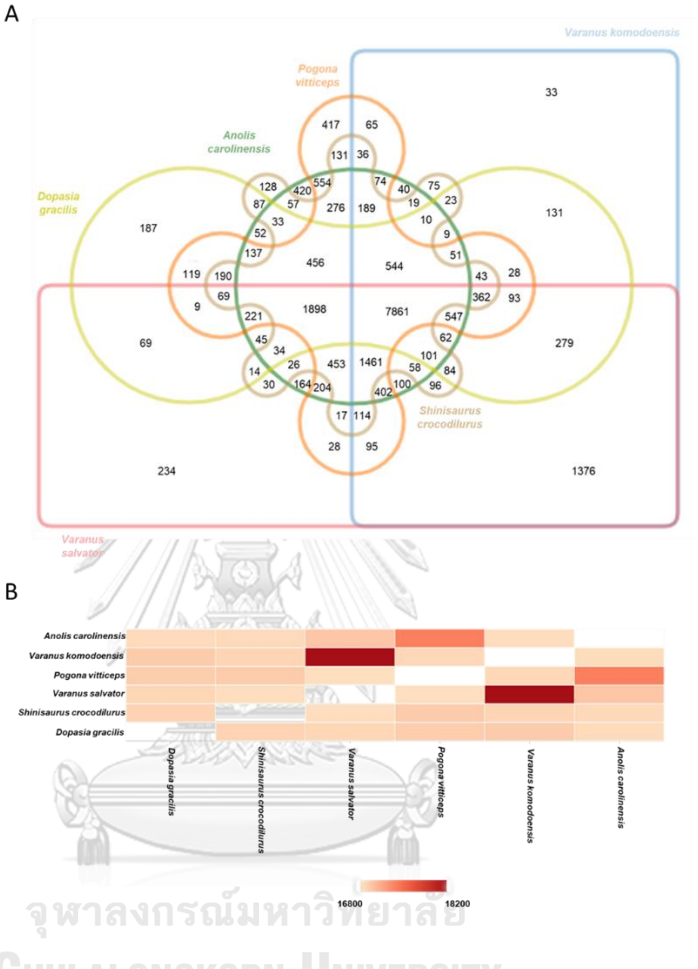

Figure 7 Gene family clustering of *V.salvator* visualized by venn diagram (A) and heatmap (B)

4.1.5 Positive selection

To determine the protein evolution in the *V. salvator* genome, we validated the positive selection with 1339 one-to-one orthologues from one the six genomes of the same clade as *V.salvator*. A gene would be considered under positive selection for the *V. salvator* lineage if a p-value was lower than 0.05. Our results found 262 genes with a positive selection signal in *V. salvator* (Supplementary Table S5). To determine the Gene Ontology (GO) and gene set enrichment analysis of the metabolic pathway. We found evidence for pathway

enrichment across the regulators of blood coagulation (p-value = 0.027) and Toll receptor signaling pathway (p-value = 0.0426). For the blood coagulation pathway genes F3, PLAU, and PLAT genes, the dN/dS ratios were 5.26, 4.19, and 9.89, respectively. Within the Toll receptor signaling pathway, which is in a family of evolutionarily strong conserved innate immune receptors, there were three genes (TAB1, CHUNK (IKK-α), and NFKBIE) with dN/dS ratios of 1.09, 1.29, and 1.53, respectively. The coagulation was part of the innate immune system which is the most critical pathway during activation of the early host response mechanism (22). Our findings support the genes controlling blood clotting and innate immunity genes, which development of a robust immune system in *V. salvator* via positive selection.

4.2 *Megaustenia siamensis*

4.2.1 De novo genome assembly of *M.siamensis*

The genome size of *M. siamensis* is estimated to be around 2.2 Gb with 0.358% heterozygosity by raw data of short-read sequencing data (Figure 8A). We obtained 234 Gb of raw PacBio long-reads. Sequences shorter than 1000bp and quality (QV) < 0.8 were filtered out. Remaining 196 Gb were used for de novo assembly with CANU software. It resulted in an assembly of 3.07 Gb containing a total of 5,246 contigs with an N50 contigs length of 1.5 Mb.

Then, we obtained 2.6 Gb, and N50 of 1.8 Mb after using Illumina short read for polished and error correction. The final draft genome assembly was performed genomic scaffolding using Hi-C data. A total of 161 scaffolds were anchored into 32 pseudo-chromosomes (2n=64) (Figure 8B) with a total length of 2.59 Gb, with an N50 of 84.3 Mb (Figure 8C). The number of chromosome scale is consistent with the land snails based on cytogenetic reports, which range from 2n=16 to 2n= 66 (40). The genome assembly sizes of *M. siamensis* is bigger than other land snails (~1.8 Gb for *Achatina fulica*, ~1.5 Gb for *Arion vulgaris*, and 1.2 Gb for *Candidula unifasciata)*. This suggests that genome size is diverse within snails (Supplementary Table S6). The GC content of *M. siamensis* is 38.24%, similar to that of the *A. vulgaris* (38.46%).

The Benchmarking Universal Single-Copy Orthologs pipeline [BUSCO v4.0.2] was used to evaluate the completeness of the genome and obtain the percentage of single-copy orthologous with metazoa_odb10 BUSCO set (954 gene metazoa gene set). 819 (85.9%) of the 954 complete expected vertebrata genes were identified as complete, including 762 (79.9%) single-copy and 57 (6.0%) duplicated. 17 (1.8%) fragmented vertebrates were present, possibly due to incomplete assembly, and only 118 (12.3%) genes were considered missing in the genome assembly. Detailed genome statistics are presented in Table 2. The completeness of the *M.siamensis* assembly was comparable to other published mollusk genome assemblies.

	Megaustenia siamensis
Genome size (bp)	2,593,626,580
Number of scaffolds	161
Scaffold N50	84,301,762
Number of protein-coding genes	34,882
Repeat content, %	60.69
GC content (%)	38.24
Complete BUSCO, %	85.9
Complete and Single-copy BUSCO (%)	79.9
Complete and Duplicated BUSCO (%)	6.0
Fragmented BUSCO (%)	1.8
Missing BUSCO (%)	12.3
Total number of metazoa_odb10	954

Table 2 Genome statistics of *M.siamensis*

BUSCO was also run with the same parameters on 13 molluscan genomes for comparative analyses (*Achatina fulica*, *Aplysia californica*, *Arion vulgaris*, *Biomphalaria glabrata*, *Candidula unifasciata*, *Chrysomallon squamiferum* (Scaly-foot gastropod), *Gigantopelta aegis*, *Haliotis rufescens* (red abolone), *Lottia gigantea*, and *Pomacea canaliculated*, *Crassostrea gigas* (Pacific oyster), *Dreissena polymorpha*, and *Mizuhopecten yessoensis*). (Figure 8D, Supplementary Table S6). The findings indicate that the genome assembly of land snails contains more duplicated genes than other groups. A high level of duplication may be explained by a whole duplication event (biological), which corresponds with Liu, Conghui, et al. report that whole genome duplication (WGD) is played an important role in the terrestrial of giant African snails (41). The results showed that the *M.siamensis* genome is sufficient as a resource in genome reference.

Figure 8 Genome size estimation (A), Hi-C linkage density histogram (B) Genome assembly (C), Comparison genome completeness (D), and Interspersed repeat landscape (E) of *M.siamensis*

4.2.2 Genome annotation

We utilized the RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) for run de novo repeat annotation and constructed species-specific repeat sequence libraries for the *M.siamensis.* We then used the

sequence libraries as a query to mask repetitive elements with RepeatMasker (http://www.repeatmasker.org). We found that repetitive elements accounted for 60.69% of the *M.siamensis* genome (Figure 8E, Supplementary Table S7). The characterization revealed Transposable elements (TEs) including Retroelements (ClassI) as the major repeat type, which contribute to ~1.45% of the genome followed by DNA transposons (ClasssII) at ~1.43% of the genome. The fraction of TEs estimated in mollusc genomes varies between 2 and 8% (42) Remaining 54.83 % is unclassified to TEs classification, which RepeatModeler's classifications are based on sequence similarity to already-known TE families in the RepeatMasker database. As a result, it is more likely to discover novel TE families

A total of 34,882 protein-coding genes were annotated in the *M.siamensis* genome. Among the predicted protein-coding genes, 58% could be annotated through at least one of the following protein-related databases: the EggNOG database (20,197:57.90%), the Swiss-Prot protein database (14,181:40.65%), the protein families (Pfam) database (18,863:54.08%), and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (11,103:31.83%).

4.2.3 Phylogenetic and comparative analysis

We constructed a phylogenetic tree to understand the genome evolution of *M.siamensis* by IQ-TREE with 1,000 bootstrap. The IQ-TREE input data was the multiple sequence alignment from the PRANK using 26 single-copy orthologous genes, which were analyzed with OrthoFinder version 2.3.12 by comparing the protein sequences of 10 gastropod species (*Achatina fulica*, *Aplysia californica*, *Arion vulgaris*, *Biomphalaria glabrata*, *Candidula unifasciata*, *Chrysomallon squamiferum* (Scaly-foot gastropod), *Gigantopelta aegis*, *Haliotis rufescens* (red abolone), *Lottia gigantea*, and *Pomacea canaliculate*) and 3 bivalvia species (*Crassostrea gigas* (Pacific oyster), *Dreissena polymorpha*, and *Mizuhopecten yessoensis*).

According to the tree, *M. siamensis* was most closely related to *A.vulgaris*, with an estimated divergence time of approximately 63 million years ago (MYA). Land snails split from other habitats (freshwater and marine snails) around 254 Mya. Gastropoda and Bivalvia split to 439 million years ago.

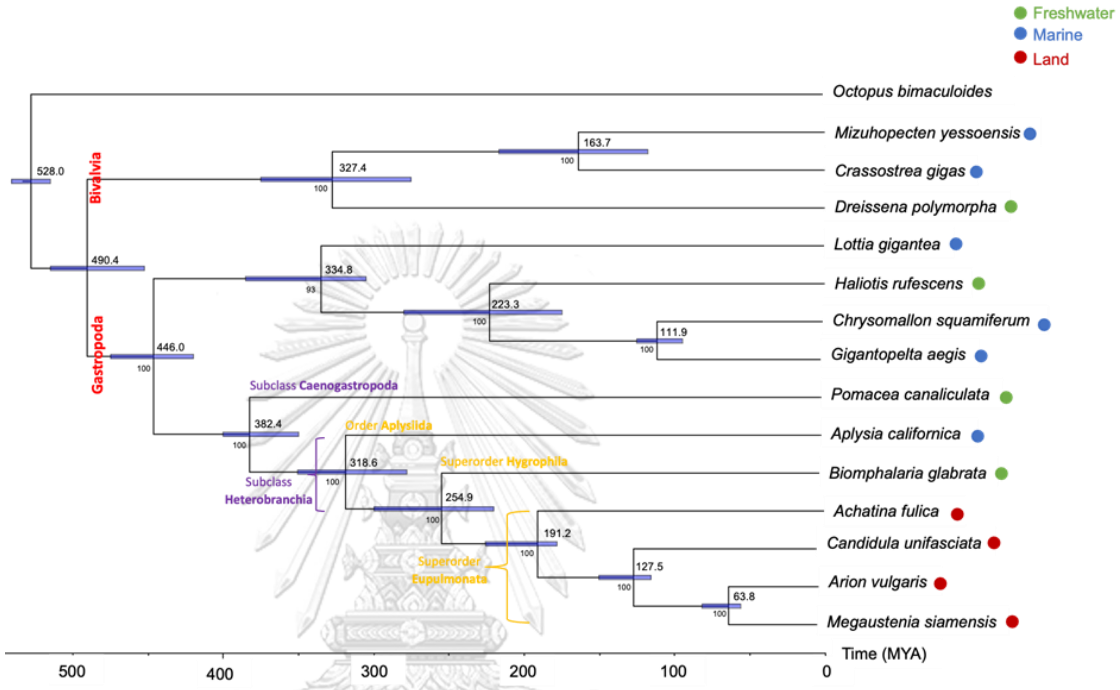

Figure 9 Phylogenetic relationship of *M.siamensis*. The divergence times [million year ago (MYA)] with 95% confidence intervals represent blue color bar.

4.2.4 Gene family clustering

We use Orthovenn2 to identify species-specific genes, presence/absence of gene families, and pathway enrichment among the four genomes of the land snails including *M.siamensis*, *A.vulgaris*, *C.unifasciata*, and *A.fulica*. The selected genome formed 18,861 clusters of 16,787 orthologous clusters (containing two species at least) and 2,074 single-copy gene clusters. The Venn diagram shows that 4,349 gene families were shared among the four genomes supporting their conservation in the lineage (Figure 10A). A total of 405 out of 4,349 gene families are specific to *M.siamensis*. We were able to annotate 117 out of 405 clusters according to the Swiss Protein Database, including 92 clusters (Supplementary Table S8) in biological processes (BP), 16 in molecular

functions (MF), and 9 in cellular components (CC). GO enrichment reveals function involve response to bacterium (GO:0009617, GO level: BP, P = 0.37e-3) (Supplementary Table S9). Figure 10B shows the pairwise heatmap based on the similarity matrix, which illustrates the overlapping cluster numbers for the four genomes. Both *M.siamensis* and *C.unifasciata* shared the highest number of orthologue clusters (18,199), followed by *A.vulgaris*. The result corresponds to the phylogenetic tree.

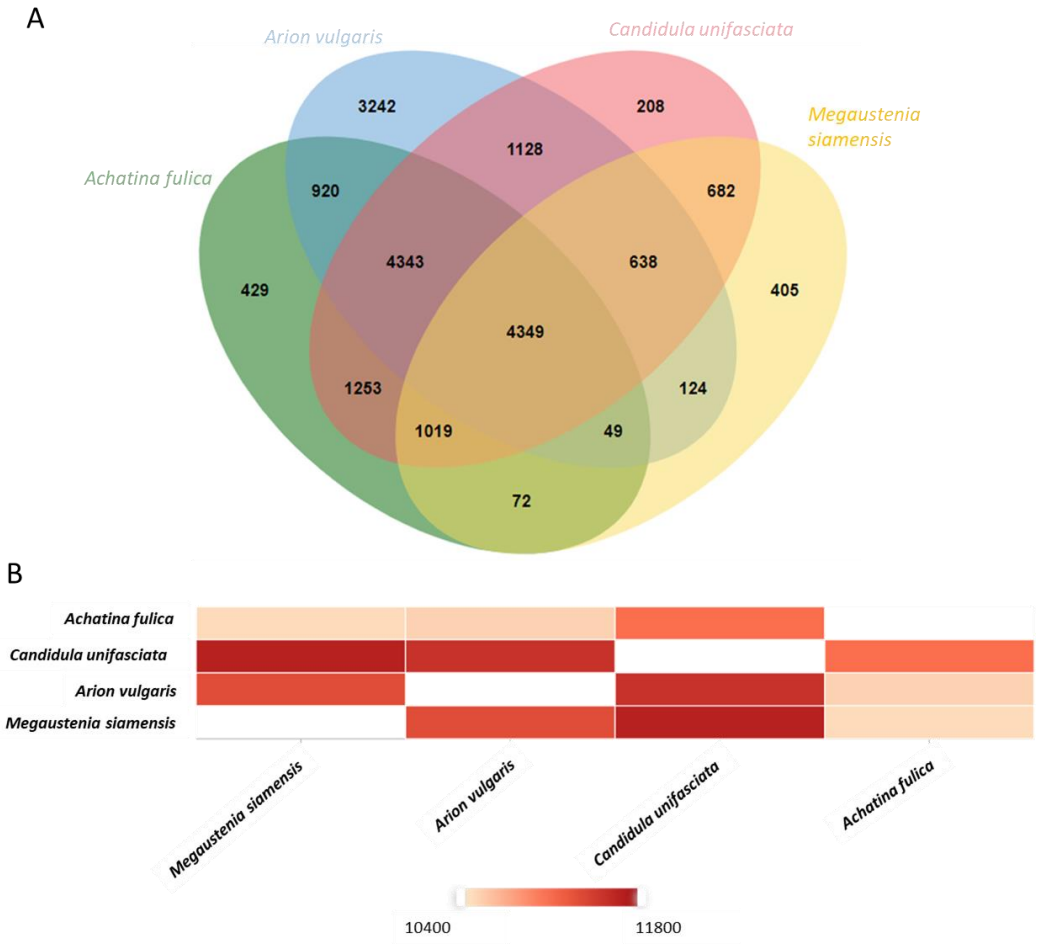

Figure 10 Gene family clustering of *M.siamensis* visualized by venn diagram (A) and heatmap (B)

4.2.5 Positive selection

We further explored the roles of positive selection genes (dn/ds > 1), we identified 17 out of 26 orthologous groups are likely positively selected in *M.siamensis* (Supplementary Table 10). The gene ontology

enrichment of genes under positive selection including 3 biological process, 7 cellular component, and 4 molecular function (Supplementary Table 6). We also found evidence for pathway enrichment across the ubiquitin proteasome pathway (PSMD1 genes) with a raw p=0.004. Our finding support the protein binding in *M.siamensis* via positive selection.

4.2.6 Transcriptome assembly and Functional annotation

The comparative transcriptome of *M.siamensis* foot and mantle were generated and assembled using Illumina Hiseq sequencing platform. Approximately 30 Gb were assembled into 716,193 transcripts with an average length of 570 bp, 403,122 unigenes, and 39.19 % of the GC content. The unigenes were annotated in the Refseq (70,286, 17.44%), Pfam (69,212, 17.17%), COG (68,487, 16.99%), GO (64,830, 16.08%), and KEGG (5,981, 1.48%). A total of 7,830 genes, encoding putative antimicrobial peptides (AMPs) in the *M.siamensis* transcriptomes.

According to GO annotation, 64,830 unigenes were categorized into 113 GO terms in the biological process, cellular component, and molecular function ontologies. In the biological process category, most of the unigenes were clustered into metabolism (4,342, 6.70%), development (3,097, 4.77%,), and cell organization and biogenesis (2,102, 3.24%). The most represented categories among the cellular component category were cell (1,667, 2.57%), intracellular (1,281, 1.98%), and cytoplasm (630, 0.97%). In the molecular function category, the matched sequences were divided into catalytic activity (2,016, 3.11%), binding (1,130, 1.74%), and transferase activity (715, 1.10%) (Figure 11A).

According to COG annotation, 68,487 unigenes were categorized into 24 functional classifications (Figure 11B). The categories with the highest proportion of unigenes were Function unknown (20,195, 29.49%), signal transduction mechanisms (8,636, 12.61%), and Transition, ribosomal structure and biogenesis (6,715,

9.80%). There were 5,981 unigenes significantly matched in the KEGG database and were assigned to 419 KEGG pathways. The pathway with the highest proportion of unigenes were Metabolic pathways followed by Biosynthesis of secondary metabolites.

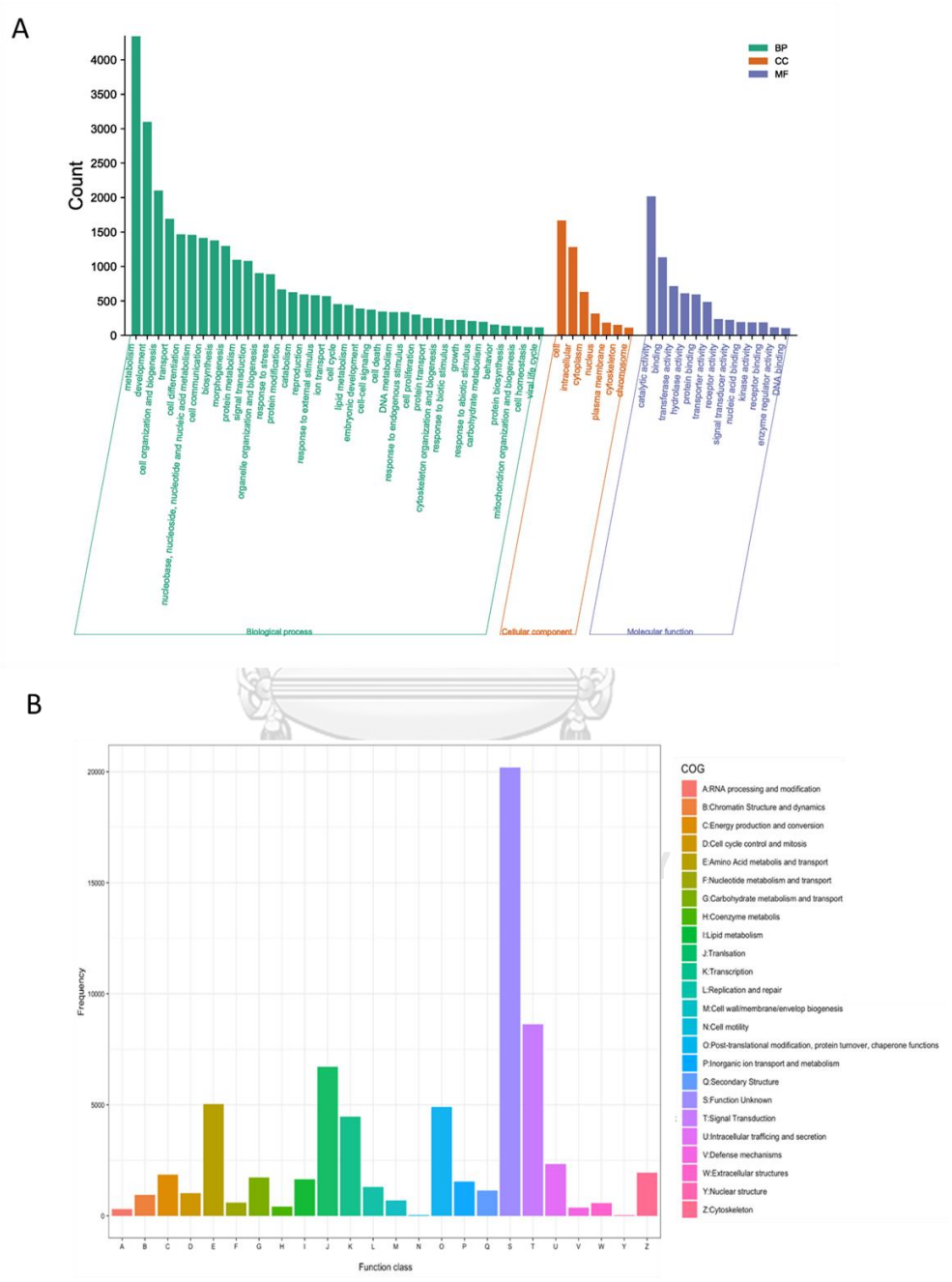

Figure 11 Function classification in Gene ontology (A) and Clusters of Orthologous Groups of proteins (COG) classification of all unigenes in the *M.siamensis* transcriptome (B)

4.2.8 Identification of Differentially expressed Genes (DEGs)

A total of 390,055 transcripts were assembled with an average length of 636 bp and 285,630 unigenes from the *M. siamensis* foot muscle data. 427,574 transcripts were assembled with an average length of 555 bp and 255,705 unigenes from the *M. siamensis* mantle data. The GC percentage of the unigenes was 38.75% in the foot muscle and 39.32% in the mantle. We detected 115,263 (55,675 up-regulated and 59,588 down-regulated) differentially expressed genes (DEGs) from the *M. siamensis* foot and mantle. The top 20 most highly expressed genes of the foot muscle and mantle are similar, as shown in Table 2 and Table 3, respectively. Most proteins responsible for biological adhesion are expressed higher in the foot than in the mantle (Figure 12), which can category into two major class, lectin-like proteins(C-lectins, C1q, and H-lectins) and matrilin-like proteins (VWA and EGF). Furthermore, actin is more abundant in the foot than in the mantle. It helps in cell adhesion and involves movement. These proteins are consistent with the mechanism of glue(43).

ID	PValue	FDR	Expected count	Description
TRINITY_GG_26624_c14_g1	1.31E-55	4.89E-50	4920960.8	
TRINITY_GG_26465_c230_g1	1.63E-51	3.04E-46	1917161.26	Actin
TRINITY_DN1068_c2_g1	5.09E-49	4.76E-44	1079062.74	
TRINITY_GG_58073_c6_g1	3.85E-48	2.88E-43	881383.73	
TRINITY_GG_2923_c17_g1	1.78E-47	1.11E-42	756239.52	
TRINITY_GG_23565_c7_g1	5.47E-47	2.92E-42	676045.46	
TRINITY_GG_3496_c104_g1	1.1E-46	4.55E-42	630522.71	
TRINITY_DN1068_c4_g1	2.9E-46	9.82E-42	572116.74	VWA
TRINITY_GG_30273_c12_g1	3.16E-46	9.82E-42	567335.7	
TRINITY_DN28462_c19_g1	3.82E-46	1.1E-41	556665.62	Myosin tail
TRINITY_DN2111_c2_g1	5.98E-46	1.6E-41	532203.59	VWA
TRINITY_GG_57043_c55_g1	8.29E-46	1.97E-41	515122.85	Tropomyosin
TRINITY_DN747_c1_g2	5.07E-45	9.97E-41	429752.04	GTP_EFTU
TRINITY_GG_17698_c0_g1	1.27E-44	2.03E-40	392139.84	Lectin_C
TRINITY_GG_40117_c18_g1	1.3E-44	2.03E-40	391077.26	
TRINITY_GG_2044_c58_g1	1.36E-44	2.03E-40	389519.44	Myosin_tail_1
TRINITY_GG_28374_c55_g1	1.36E-44	2.03E-40	389360.14	EF-hand_6
TRINITY_GG_8818_c120_g1	2.74E-44	3.42E-40	362997.33	
TRINITY_GG_49898_c29_g1	6.31E-44	6.93E-40	333985.07	VWA

Table 3 The top 20 Most highly expressed genes in the Foot transcriptome.

ID	PValue	FDR	Expected count	Description
TRINITY_GG_27001_c17_g1	7.61E-50	9.47E-45	1184032.67	
TRINITY_GG_2568_c1348_g1	8.05E-47	3.76E-42	590117.22	
TRINITY_DN1503_c55_g1	2.03E-46	7.59E-42	537900.98	
TRINITY_DN1830_c0_g1	1.11E-43	1.15E-39	504381.52	
TRINITY_GG_1231_c6_g1	8.43E-46	1.97E-41	466578.89	
TRINITY_GG_2568_c1_g1	2.11E-45	4.63E-41	425713.02	
TRINITY_GG_11003_c29_g1	2.33E-45	4.83E-41	421511.74	
TRINITY_GG_2568_c9_g1	6.05E-45	1.13E-40	383078.98	
TRINITY_GG_25287_c221_g1	1.32E-44	2.03E-40	354402.4	Actin
TRINITY_GG_8829_c14_g1	1.85E-44	2.66E-40	342572.65	
TRINITY_DN610_c0_g2	1.96E-44	2.71E-40	340640	Kunitz_BPTI
TRINITY_GG_2568_c18_g1	2.13E-44	2.84E-40	337791.24	Chitin binding Peritrophin-A domain
TRINITY_GG_24988_c45_g1	2.49E-44	3.2E-40	332583.65	
TRINITY_GG_45343_c185_g1	2.92E-44	3.52E-40	327313.8	
TRINITY_DN2216_c0_g2	3.8E-44	4.43E-40	318838.31	VWA
TRINITY_GG_25019_c3_g1	4.34E-44	4.92E-40	314564.03	
TRINITY_GG_47492_c21_g1	7.12E-44	7.6E-40	299396.41	
TRINITY_DN1_c45_g2	3.47E-43	3.09E-39	255510.02	GTP_EFTU
TRINITY_GG_8506_c29_g2	4.42E-43	3.75E-39	249444.38	
TRINITY_DN3375_c0_g1	5.88E-27	1.32E-24	247746.27	HMG_box_2

Table 4 The top 20 Most highly expressed genes in the Mantle transcriptome.

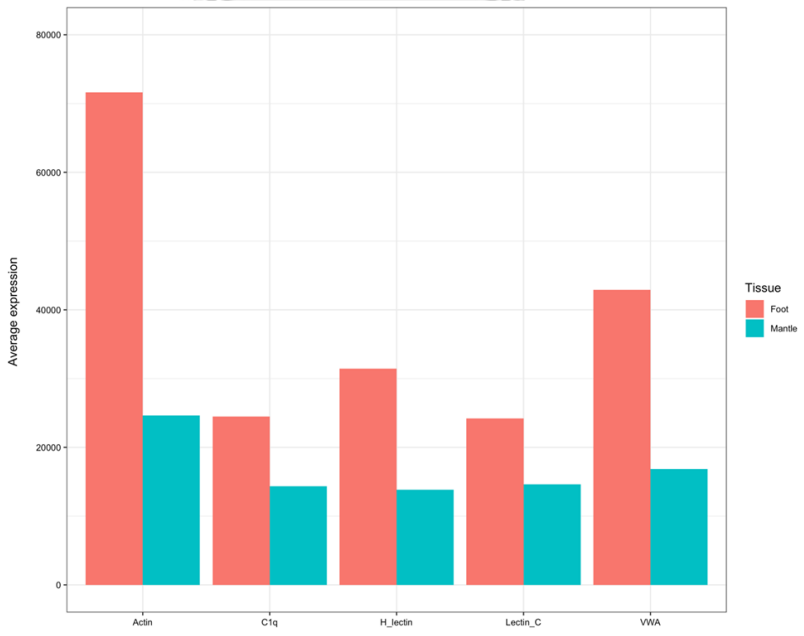

Figure 12 Protein expression involving the glue mechanism.

4.2.9 Antimicrobial and anticancer activity prediction

A total of 7,830 sequences were encoded putative antimicrobial peptides (AMPs) in the *M.siamensis* transcriptomes. Then, sequences were analyzed in CAMP database by the four machine-learning algorithms, such as Support Vector Machine (SVM), Discriminant Analysis (DA), Artificial Neural Network (ANN), and Random Forest (RF) and iACP tool for anticancer. It resulted in 44 putative active peptides. Of these, 29 sequences were predicted to be only antimicrobial, Bacteriocin families are the most identified peptides. 8 sequences were both putative antimicrobial and anticancer and 7 sequences were predicted to be only anticancer (Figure 13). Furthermore, 8 potential active peptides are highly express in foot (4 peptides) and mantle (4 peptides). Several antimicrobial peptides from snails have been investigated (44, 45). Our results could be considered as potential alternatives in therapy.

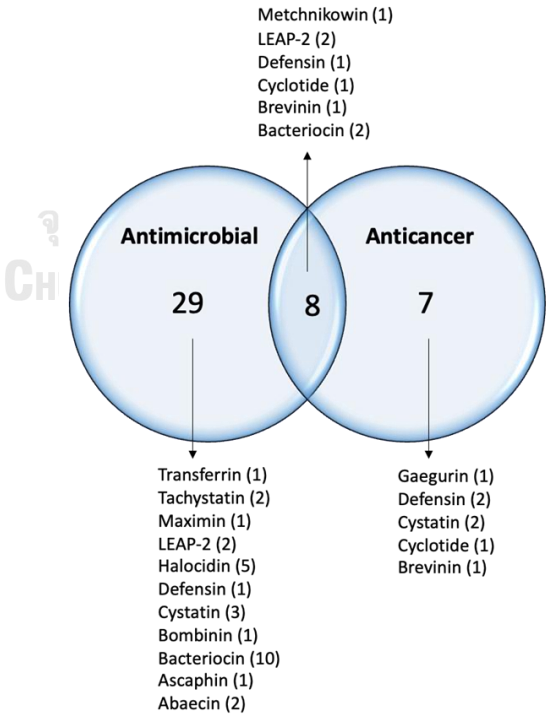

Figure 13 44 putative active peptides of antimicrobial and anticancer prediction in *M.siamensis* transcriptome

Discussion and conclusions

High throughput DNA sequencing (NGS) is used to determine the order of nucleotides in entire genomes or targeted regions of DNA or RNA, which is essential for unraveling biological data mysteries. These platforms are useful in a variety of omic-research applications, including de novo sequencing or RNA sequencing. In this study we use 2 key elements for understanding the animal application. 1) Genomics, to get a deeper understanding the evolution and species-specific traits. 2) Transcriptome, to identifying transcript and understanding which genes and their levels of expression are expressed different among tissue. However, the complete and good quality of genome assembly depend on the nature of the genome and the selection of the sequencing platform. The missing or duplicated of the assembly may be explained by a recent whole duplication event (biological) or a chimeric assembly of haplotypes (technical). Non-gapped assembly length is a good measure of the completeness of the assembly, which lead to the actual genome size of species. *Varanus salvator* using 2 sequencing technology, 10X genomics linked read and long-read pacbio. The genome completeness of *V.salvator* have more duplication than within species. This demonstrates that our *V. salvator* genome have some limitation according to our study's de novo assembly pipeline. To obtain actual genome size and accurate genome assembly, we can add more software that removes haplotigs or contigs overlap, such as purge dups. For *M.siamensis*, using 3 technology , short-read illumina, long-read pacbio, and. Hi-C data. However, The genome completeness of *M.siamensis* have high level of duplication. The nature of land snails is highly duplicated. Then, the high proportion of duplicates in *M.siamensis* depends on the nature of species. We can add whole genome duplication analysis in further study. However, two species are sufficient as an essential resource in genome reference. Transcriptome analysis revealed many novels transcribed is unknown proteins. The expression levels of an unknown gene are significantly higher in the foot than in the mantle, suggesting that the unknown gene may play a role in foot movement. The newly transcribed genes provide researchers with a good starting point for investigating the function of a newly discovered gene.

The draft genome of *V.salvator* and *M.siamensis* can serve as a reference for future research and will help to understand evolutionary adaptations. Furthermore, the transcriptome data of *M.siamensis* reveal the highest protein composition and bioactive compounds, which might be useful for medical applications.

REFERENCES

1. Qin Y, Koehler S, Zhao S, Mai R, Liu Z, Lu H, et al. High-throughput, low-cost and rapid DNA sequencing using surface-coating techniques. bioRxiv. 2020.

2. Kulabtong S, and Rujira Mahaprom. Observation on food items of Asian water monitor, Varanus salvator (Laurenti, 1768)(Squamata Varanidae), in urban eco-system, Central Thailand. Biodiversity Journal 2015.

3. Liao X, Li M, Zou Y, Wu F-X, Yi P, Wang J. Current challenges and solutions of de novo assembly. Quantitative Biology. 2019;7(2):90-109.

4. Jung H, Ventura T, Chung JS, Kim WJ, Nam BH, Kong HJ, et al. Twelve quick steps for genome assembly and annotation in the classroom. PLoS Comput Biol. 2020;16(11):e1008325.

5. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13(5):278-89.

6. Guiglielmoni N, Rivera-Vicéns R, Koszul R, Flot J-F. A deep dive into genome assemblies of non-vertebrate animals. Peer Community Journal. 2022;2.

7. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33(14):2202-4.

8. Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Vinnere Pettersson O, et al. Ten steps to get started in Genome Assembly and Annotation. F1000Res. 2018;7.

9. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5.

10. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol. 2019;1962:227-45.

11. Ejigu GF, Jung J. Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. Biology (Basel). 2020;9(9).

12. Koonin EV, Galperin MY. Comparative Genomics and New Evolutionary Biology. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston2003.

13. Altenhoff AM, Glover NM, Dessimoz C. Inferring Orthology and Paralogy. Methods Mol Biol. 2019;1910:149-75.

14. Wagner A. Rapid detection of positive selection in genes and genomes through variation clusters. Genetics. 2007;176(4):2451-63.

15. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12(10):671-82.

16. Yang IS, Kim S. Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. Genomics Inform. 2015;13(4):119-25.

17. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 2014;15(12):553.

18. Shine R, Harlow PS, Keogh JS. Commercial harvesting of giant lizards: the biology of water monitors Varanus salvator in southern Sumatra. Biological Conservation. 1996;77(2-3):125-34.

19. Akbar N, Siddiqui R, Sagathevan K, Iqbal M, Khan NA. Gut Bacteria of Water Monitor Lizard (Varanus salvator) Are a Potential Source of Antibacterial Compound(s). Antibiotics (Basel). 2019;8(4).

20. Rios FM, and Laura M. Zimmerman. Immunology of reptiles. eLS. 2015:1-7.

21. Lind AL, Lai YYY, Mostovoy Y, Holloway AK, Iannucci A, Mak ACY, et al. Genome of the Komodo dragon reveals adaptations in the cardiovascular and chemosensory systems of monitor lizards. Nat Ecol Evol. 2019;3(8):1241-52.

22. van der Poll T, Herwald H. The coagulation system and its function in early immune defense. Thromb Haemost. 2014;112(4):640-8.

23. Gazquez-Gutierrez A, Witteveldt J, S RH, Macias S. Sensing of transposable elements by the antiviral innate immune system. RNA. 2021.

24. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722-36.

25. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 2016;44(19):e147.

26. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

27. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963.

28. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.

29. Loytynoja A. Phylogeny-aware alignment with PRANK. Methods Mol Biol. 2014;1079:155-70.

30. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268-74.

31. Xu L, Dong Z, Fang L, Luo Y, Wei Z, Guo H, et al. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 2019;47(W1):W52-W8.

32. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34(Web Server issue):W609-12.

33. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res. 2003;31(1):334-41.

34. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494-512.

35. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-9.

36. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

37. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.

38. Almagro Armenteros JJ, Tsirigos KD, Sonderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019;37(4):420-3.

39. Waghu FH, Idicula-Thomas S. Collection of antimicrobial peptides database and its derivatives: Applications and beyond. Protein Sci. 2020;29(1):36-42.

40. Khrueanet W, Weerayuth Supiwong, Chanidaporn Tumpeesuwan, Sakboworn Tumpeesuwan, Krit Pinthong, and Alongklod Tanomtong. First chromosome analysis and localization of the nucleolar organizer region of land snail, Sarika resplendens (Stylommatophora, Ariophantidae) in Thailand. Cytologia. 2013;78:213-22.

41. Liu C, Ren Y, Li Z, Hu Q, Yin L, Wang H, et al. Giant African snail genomes provide insights into molluscan whole-genome duplication and aquatic-terrestrial transition. Mol Ecol Resour. 2021;21(2):478-94.

42. Thomas-Bulle C, Piednoel M, Donnart T, Filee J, Jollivet D, Bonnivard E. Mollusc genomes reveal variability in patterns of LTR-retrotransposons dynamics. BMC Genomics. 2018;19(1):821.

43. Smith AM, Papaleo C, Reid CW, Bliss JM. RNA-Seq reveals a central role for lectin, C1q and von Willebrand factor A domains in the defensive glue of a terrestrial slug. Biofouling. 2017;33(9):741-54.

44. Cilia G, Fratini F. Antimicrobial properties of terrestrial snail and slug mucus. J Complement Integr Med. 2018;15(3).

45. Noothuan N, Apitanyasai K, Panha S, Tassanakajon A. Snail mucus from the mantle and foot of two land snails, Lissachatina fulica and Hemiplecta distincta, exhibits different protein profile and biological activity. BMC Res Notes. 2021;14(1):138.

APPENDIX

Supplementary Table S1. Genome statistics and sources for non-avian reptiles used in this study.

Species	Species common name	Assembly size (Gb)	GC content (%)	Protein	Source	Accession
Alligator mississippiensis	American alligator	2.16	44.3	42,388	NCBI	GCF_000281125.3
Alligator sinensis	Chinese alligator	2.27	44.6	43,092	NCBI	GCF_000455745.1
Anolis carolinensis	Green anole lizard	1.79	40.82	34,827	NCBI	GCF_000090745.1
Chelonia mydas	Green sea turtle	2.2	43.7	14,342	NCBI	GCF_000344595.1
Chrysemys picta	Western painted turtle	2.33	44.2	46,651	NCBI	GCF_000241765.3
Crocodylus porosus	Saltwater crocodile	2.05	43.85	28,676	NCBI	GCF_001723895.1
Deinagkistrodon acutus	Five-pacer viper	1.47	-	21194	GigaDB	DOI: 10.5524/100196
Dopasia gracilis	Burmese glass lizard	1.78	43.71	19,513	GigaDB	DOI: 10.5524/100119
Eublepharis macularius	Leopard gecko	2.02	43.55	24,755	GigaDB	DOI: 10.5524/100246
Gavialis gangeticus	Indian gharial	2.64	44.95	27,294	NCBI	GCF_001723915.1
Gekko japonicus	Japanese gecko	2.49	45.5	24,464	NCBI	GCF_001447785.1
Ophiophagus hannah	King cobra	1.59	40.6	18,445	NCBI	GCA_000516915.1
Pelodiscus sinensis	Chinese Soft-shell turtle	2.2	44.49	38,587	NCBI	GCF_000230535.1
Pogona vitticeps	Australian dragon lizard	1.71	42.1	38,712	NCBI	GCF_900067755.1
Protobothrops mucrosquamatus	Pit viper	1.67	40.6	23,351	NCBI	GCF_001527695.2
Python bivittatus	Burmese python	1.43	39.7	32,603	NCBI	GCA_000186305.2
Shinisaurus crocodilurus	Chinese crocodile lizard	2.24	44.46	20,150	GigaDB	DOI: 10.5524/100315
Varanus komodoensis	Komodo dragon	1.51	44.04	18,293	NCBI	GCA_004798865.1
Varanus salvator macromaculatus	**Water monitor**	**1.7**	**44.03**	**21,937**	**This study**	**This study**

Supplementary Table S2. Repetitive elements in *Varanus salvator*

sequences: 858 total length: 1702614977 bp GC level: 44.03 % bases masked: 637321099 bp (37.43 %)				
Type	SubType	No.of Element	Length (bp)	Proportion of Genome (%)
SINE		180868	180868	1.66
	ALUs	0	0	0.00
	MIRs	156619	26380246	1.55
LINEs		755276	285317750	16.76
	LINE1	40529	19350788	1.14
	LINE2	283819	85305588	5.01
	L3/CR1	277002	103430321	6.07
LTR elements		33905	28040502	1.78
	ERVL	540	34026	0.00
	ERVL-MaLRs	76	4318	0.00
	ERV_classI	12306	1639513	0.10
	ERV_classII	3528	593064	0.04
DNA elements		154136	23375942	1.48
	hAT-Charlie	30207	6384221	0.40
	TcMar-Tigger	16165	3140236	0.20
Unclassified		1353967	227439741	14.42
Total interspersed repeats			567698611	35.99
Small RNA		35088	3460605	0.22
Satellites		3036	259855	0.02
Simple repeats		292591	10398879	0.66
Low complexity		29001	1290060	0.08

Supplementary Table S3. 94 unique gene clusters specific in *V. salvator*

Cluster_name	Protein_number	Swiss_prot_id	Go_annotation
cluster21095	2	Q99626	GO:0001829; P:trophectodermal cell differentiation; IEA:Ensembl
cluster21131	2	Q00534	GO:0003323; P:type B pancreatic cell development; IDA:UniProtKB
cluster21096	2	Q8N1T3	GO:0003774; F:motor activity; IEA:InterPro
cluster875	15	P16423	GO:0003964; F:RNA-directed DNA polymerase activity; IEA:UniProtKB-KW
cluster21201	2	P18993	GO:0005344; F:oxygen carrier activity; IEA:UniProtKB-KW
cluster21204	2	P0C0U6	GO:0005344; F:oxygen carrier activity; IEA:UniProtKB-KW
cluster21107	2	P52786	GO:0005506; F:iron ion binding; IEA:InterPro
cluster21100	2	Q8C088	GO:0005509; F:calcium ion binding; IEA:InterPro
cluster21126	2	Q62082	GO:0005509; F:calcium ion binding; IEA:InterPro
cluster21173	2	Q41420	GO:0005509; F:calcium ion binding; IEA:InterPro
cluster3993	9	Q9NUV9	GO:0005525; F:GTP binding; IBA:GO_Central
cluster21054	2	Q96F15	GO:0005525; F:GTP binding; IBA:GO_Central
cluster21111	2	Q8WWP7	GO:0005525; F:GTP binding; IBA:GO_Central
cluster21189	2	Q8K3K9	GO:0005525; F:GTP binding; IBA:GO_Central
cluster5575	8	O02751	GO:0005634; C:nucleus; IEA:UniProtKB-SubCell
cluster21036	2	P0CT41	GO:0006310; P:DNA recombination; IEA:UniProtKB-KW
cluster21108	2	Q9NBX4	GO:0006313; P:transposition, DNA-mediated; IMP:UniProtKB
cluster21142	2	Q95SX7	GO:0006313; P:transposition, DNA-mediated; IMP:UniProtKB
cluster7917	7	P84227	GO:0006334; P:nucleosome assembly; IBA:GO_Central
cluster21056	2	Q8CAV0	GO:0006364; P:rRNA processing; IEA:InterPro
cluster17938	3	E1BB52	GO:0006368; P:transcription elongation from RNA polymerase II promoter; IBA:GO_Central
cluster21106	2	Q5ZLP8	GO:0006417; P:regulation of translation; IEA:UniProtKB-KW
cluster21125	2	Q9NRH2	GO:0006468; P:protein phosphorylation; IDA:UniProtKB
cluster17934	3	P51589	GO:0006805; P:xenobiotic metabolic process; IBA:GO_Central
cluster21073	2	O54749	GO:0006805; P:xenobiotic metabolic process; IBA:GO_Central
cluster21165	2	P10632	GO:0006805; P:xenobiotic metabolic process; IBA:GO_Central
cluster21058	2	P22897	GO:0006898; P:receptor-mediated endocytosis; IDA:UniProtKB
cluster21038	2	Q8WZB3	GO:0006941; P:striated muscle contraction; TAS:UniProtKB
cluster7914	7	P15979	GO:0006955; P:immune response; IBA:GO_Central
cluster17936	3	P18469	GO:0006955; P:immune response; IEA:InterPro
cluster21084	2	P01915	GO:0006955; P:immune response; IEA:InterPro
cluster21110	2	P20756	GO:0006955; P:immune response; IEA:InterPro
cluster21075	2	Q5R495	GO:0006979; P:response to oxidative stress; IEA:Ensembl
cluster21103	2	A6NMZ7	GO:0007155; P:cell adhesion; IEA:UniProtKB-KW

cluster21157	2	Q7SZ59	GO:0007165; P:signal transduction; IEA:InterPro
cluster21198	2	A6NI28	GO:0007165; P:signal transduction; IEA:InterPro
cluster21203	2	Q14123	GO:0007165; P:signal transduction; IEA:InterPro
cluster21076	2	Q14571	GO:0007165; P:signal transduction; TAS:ProtInc
cluster21062	2	O42603	GO:0007166; P:cell surface receptor signaling pathway; IEA:InterPro
cluster21140	2	P48058	GO:0007215; P:glutamate receptor signaling pathway; TAS:ProtInc
cluster21214	2	Q5BKN4	GO:0007224; P:smoothened signaling pathway; ISS:UniProtKB
cluster19609	2	P14837	GO:0007275; P:multicellular organism development; IEA:UniProtKB-KW
cluster21072	2	O75718	GO:0007283; P:spermatogenesis; IEA:Ensembl
cluster21116	2	Q9BW62	GO:0007283; P:spermatogenesis; IEA:UniProtKB-UniRule
cluster21086	2	Q3KNS1	GO:0007286; P:spermatid development; NAS:UniProtKB
cluster21199	2	Q5ZL42	GO:0007399; P:nervous system development; IEA:UniProtKB-KW
cluster21177	2	Q53QF6	GO:0007517; P:muscle organ development; TAS:ProtInc
cluster14669	5	Q5JZQ7	GO:0007601; P:visual perception; IEA:Ensembl
cluster21143	2	Q8NFP9	GO:0008104; P:protein localization; IEA:Ensembl
cluster21178	2	Q9UPY6	GO:0008360; P:regulation of cell shape; IMP:UniProtKB
cluster21200	2	P30372	GO:0009615; P:response to virus; IEA:Ensembl
cluster21162	2	Q86UI5	GO:0009617; P:response to bacterium; IEA:Ensembl
cluster21101	2	Q8BSD4	GO:0014717; P:regulation of satellite cell activation involved in skeletal muscle regeneration; IMP:MGI
cluster21179	2	Q68F72	GO:0015031; P:protein transport; IEA:UniProtKB-KW
cluster21145	2	A4IH68	GO:0015937; P:coenzyme A biosynthetic process; IBA:GO_Central
cluster21077	2	Q8N2W9	GO:0016055; P:Wnt signaling pathway; IEA:UniProtKB-KW
cluster14671	5	Q9C029	GO:0016567; P:protein ubiquitination; IEA:UniProtKB-UniPathway
cluster21089	2	Q5ZJD8	GO:0016746; F:transferase activity, transferring acyl groups; IEA:InterPro
cluster5574	8	Q6TAC4	GO:0019236; P:response to pheromone; IEA:UniProtKB-KW
cluster7915	7	E9Q6I0	GO:0019236; P:response to pheromone; IEA:UniProtKB-KW
cluster21171	2	Q10835	GO:0022617; P:extracellular matrix disassembly; ISS:UniProtKB
cluster21050	2	Q9UPZ6	GO:0030154; P:cell differentiation; IEA:UniProtKB-KW
cluster21151	2	Q28397	GO:0030574; P:collagen catabolic process; IEA:UniProtKB-KW
cluster21081	2	Q8BHD0	GO:0032482; P:Rab protein signal transduction; IBA:GO_Central
cluster21048	2	Q9NZ43	GO:0032940; P:secretion by cell; ISS:HGNC
cluster19833	2	Q9UBL0	GO:0034605; P:cellular response to heat; IEA:Ensembl
cluster21195	2	O15162	GO:0035456; P:response to interferon-beta; IMP:UniProtKB
cluster17939	3	P13761	GO:0042088; P:T-helper 1 type immune response; IMP:UniProtKB
cluster5576	8	P28068	GO:0042102; P:positive regulation of T cell proliferation; IMP:UniProtKB
cluster21161	2	Q62158	GO:0042147; P:retrograde transport, endosome to Golgi; ISS:UniProtKB
cluster21192	2	P59111	GO:0042391; P:regulation of membrane potential; IBA:GO_Central
cluster21193	2	P54277	GO:0042493; P:response to drug; IEA:Ensembl

cluster21074	2	Q9UG22	GO:0042802; F:identical protein binding; IPI:IntAct
cluster21170	2	Q9Y6M1	GO:0043488; P:regulation of mRNA stability; TAS:Reactome
cluster14672	5	P03364	GO:0044826; P:viral genome integration into host DNA; IEA:UniProtKB-KW
cluster21137	2	Q5SXI5	GO:0045893; P:positive regulation of transcription, DNA-templated; IDA:UniProtKB
cluster21141	2	Q96IT1	GO:0045893; P:positive regulation of transcription, DNA-templated; ISS:UniProtKB
cluster14674	5	Q92670	GO:0046872; F:metal ion binding; IEA:UniProtKB-KW
cluster21113	2	Q96LW9	GO:0046872; F:metal ion binding; IEA:UniProtKB-KW
cluster21164	2	O14576	GO:0047496; P:vesicle transport along microtubule; IMP:UniProtKB
cluster21052	2	Q8AX98	GO:0050793; P:regulation of developmental process; IMP:UniProtKB
cluster21123	2	P18892	GO:0050852; P:T cell receptor signaling pathway; IBA:GO_Central
cluster17944	3	Q861H8	GO:0050852; P:T cell receptor signaling pathway; TAS:Reactome
cluster21124	2	Q9BCP2	GO:0050852; P:T cell receptor signaling pathway; TAS:Reactome
cluster21188	2	P05538	GO:0050852; P:T cell receptor signaling pathway; TAS:Reactome
cluster21150	2	Q49HI0	GO:0050916; P:sensory perception of sweet taste; IBA:GO_Central
cluster21138	2	Q9Z0R7	GO:0050916; P:sensory perception of sweet taste; IDA:UniProtKB
cluster14670	5	Q9Z0R8	GO:0050917; P:sensory perception of umami taste; IDA:UniProtKB
cluster21097	2	Q99PG6	GO:0050917; P:sensory perception of umami taste; ISO:MGI
cluster21118	2	O62714	GO:0051924; P:regulation of calcium ion transport; IBA:GO_Central
cluster21037	2	Q96A23	GO:0071277; P:cellular response to calcium ion; IBA:GO_Central
cluster21046	2	A6QQZ7	GO:0071896; P:protein localization to adherens junction; IEA:Ensembl
cluster21082	2	Q62765	GO:0072553; P:terminal button organization; IMP:BHF-UCL
cluster21066	2	Q13332	GO:0099560; P:synaptic membrane adhesion; IDA:SynGO

Supplementary Table S4. Function enrichment of specific gene families in *V. salvator*

GO_ID	Name	Namespace	Count	p-value
GO:0019236	response to pheromone	biological_process	12	1.37E-11
GO:0005634	nucleus	cellular_component	9	1.93E-08
GO:0006955	immune response	biological_process	6	3.89E-09
GO:0003964	RNA-directed DNA polymerase activity	molecular_function	4	5.29E-08
GO:0006313	transposition, DNA-mediated	biological_process	4	7.25E-07
GO:0005525	GTP binding	molecular_function	4	3.38E-06
GO:0050852	T cell receptor signaling pathway	biological_process	4	0.000410411
GO:0006805	xenobiotic metabolic process	biological_process	4	0.000715649
GO:0005509	calcium ion binding	molecular_function	3	0.005711783
GO:0050916	sensory perception of sweet taste	biological_process	2	0.000236354
GO:0050917	sensory perception of umami taste	biological_process	2	0.001075715
GO:0005344	oxygen transporter activity	molecular_function	2	0.001374432
GO:0071277	cellular response to calcium ion	biological_process	2	0.002904287

Supplementary Table S5 262 genes with signatures of positive selection in *V. salvator*

Orthogroup	dN/dS	p-value	Human ortholog
OG0006468	2.02	0	ACBD3
OG0007834	1.31	0.00001	ACSL4
OG0007240	1.16	0.00001	ADD1
OG0007765	1.7	0.00067	AFAP1L2
OG0006404	1.02	0	AL592490.1
OG0006569	2.04	0.00002	ALDH7A1
OG0007654	1.7	0	ANKRD11
OG0006716	1.58	0.00283	ANKRD55
OG0006146	2.68	0	ANXA6
OG0007530	3.81	0.0002	AP5M1
OG0006224	1.25	0.01501	APCDD1
OG0006014	3.1	0.00952	APIP
OG0007105	3.02	0	APOB
OG0007514	2.26	0.00049	ARHGAP18
OG0007193	2.38	0.00001	ARHGEF16
OG0007026	1.69	0	ARID4A
OG0006980	3.24	0	ARSK
OG0007611	1.08	0.03593	ASB2
OG0006775	1.63	0.00102	ASIC5
OG0006433	1.15	0.02483	ATF3
OG0006727	2.71	0.00002	ATG4C
OG0006382	1.52	0.00204	ATP10A
OG0006259	1.2	0.00269	ATP6AP2
OG0007974	1.04	0	ATP7B
OG0006379	13.11	0	B3GAT2
OG0007363	2.44	0	B4GALT4
OG0006397	1.17	0.00068	BAG3
OG0006506	2.61	0	BBS10
OG0007322	4.01	0.00315	BCCIP
OG0006738	2.52	0.03255	BET1L
OG0006313	1.66	0.00001	BNIP3
OG0008270	3.28	0	C10orf71
OG0006118	2.57	0.00005	C1GALT1C1
OG0007518	1.83	0.00024	C1orf131
OG0006753	4.01	0	C1orf198
OG0007841	1.96	0.02322	C1orf216

OG0006385	4.57	0.00792	C3orf14
OG0006986	1.42	0	C5orf22
OG0006331	1.01	0.00752	C8A
OG0006791	2.58	0	CACUL1
OG0007526	1.29	0	CC2D2A
OG0006943	2.3	0.03627	CCDC59
OG0006595	1.03	0.00008	CCDC80
OG0008162	1.07	0.00001	CCNO
OG0007553	4.45	0	CCNT2
OG0007127	2.81	0	CD44
OG0007919	1.95	0	CDON
OG0006117	6.862	0.00007	CEL
OG0007704	2.21	0.00712	CEP55
OG0007453	1.54	0	CEP78
OG0006625	1.46	0.0001	CFAP36
OG0007593	4.04	0.00225	CFAP97
OG0006873	2.58	0	CHAF1A
OG0008330	5.08	0	CHD6
OG0007174	1.29	0.00935	CHUNK
OG0007803	3.65	0.00012	CLDN10
OG0007415	1.54	0	CNTN2
OG0008101	1.24	0.00021	COG1
OG0008314	1.31	0.00018	COG3
OG0006247	1.04	0.00253	COLEC10
OG0007588	4.38	0	COMMD8
OG0007307	2.48	0.00201	COQ4
OG0007964	2.05	0.00002	CRB1
OG0007358	2.88	0.00251	CREG1
OG0006544	3.28	0.00046	CRISPLD1
OG0006991	3.84	0	CSF1R
OG0007222	1.39	0.02002	CYP7B1
OG0006223	1.7	0	CYTIP
OG0006975	1.26	0.00003	D2HGDH
OG0006350	1.65	0	DCAF17
OG0008150	2.5	0.00001	DCSTAMP
OG0006155	3.91	0	DERA
OG0006240	1.69	0.00003	DHX33
OG0006380	1.83	0	DIPK1C
OG0007538	3.62	0	DLGAP5

OG0007928	1.91	0	DNALI1
OG0007057	1.38	0	DNTTIP2
OG0007953	1.13	0	DPH2
OG0006050	1.29	0.00004	EIF2S2
OG0006071	1.28	0.01831	ELMOD1
OG0007914	1.62	0.00015	ELMOD2
OG0008113	2.15	0	EPN2
OG0006842	5.26	0	F3
OG0006098	3.86	0.01487	FABP2
OG0007987	2.03	0	FAM124B
OG0006921	2.34	0	FAM149A
OG0008201	1.15	0	FAM83B
OG0007210	1.41	0	FBXL7
OG0008085	3.47	0.01382	FBXO2
OG0007533	1.42	0.00002	FBXO34
OG0007830	4.01	0	FLAD1
OG0007493	1.94	0.00002	FRRS1
OG0006342	1.96	0.00011	FUCA2
OG0006310	3.05	0.00009	FUT7
OG0007690	1.48	0.00086	FZD3
OG0007613	1.92	0.00064	GALC
OG0007605	2.1	0	GATB
OG0006250	2.16	0.0001	GDF9
OG0007229	2.95	0.00003	GNL2
OG0007031	2.74	0.0001	GPA33
OG0007304	1.05	0.00001	GPATCH3
OG0006548	1.48	0.00004	GPRC5C
OG0006457	6.14	0.00731	GSC
OG0006179	2.26	0.00192	GTF3C6
OG0008275	3.1	0.00009	GXYLT2
OG0007836	1.55	0.00039	HDX
OG0007519	1.45	0	HEATR1
OG0006529	1.55	0	HEXD
OG0006810	1.79	0.00001	HMMR
OG0006041	6.64	0.00454	HPGD
OG0006664	1.36	0.00007	HSF2
OG0006579	1.29	0.00415	HSPB2-C11orf52
OG0006636	1.25	0.00001	IARS2
OG0006492	4.27	0	IFNGR1

OG0007930	3.69	0	IL10RA
OG0008141	1.7	0	IL17RD
OG0007807	1.57	0	INSYN2B
OG0008071	1.86	0.0461	IQCD
OG0006299	1.94	0	KBTBD12
OG0006343	1.49	0.0246	KBTBD8
OG0007843	2.9	0	KCNG4
OG0006818	1.23	0.00011	KCTD9
OG0007211	1.65	0.00559	KIAA0930
OG0008026	1.65	0.00001	KLF11
OG0007545	1.65	0	LCT
OG0008179	5.3	0	LGSN
OG0008279	1.13	0.00112	LMCD1
OG0007497	4.55	0.00004	LRPAP1
OG0007426	3.41	0.04201	LRRC28
OG0006271	1.29	0.00004	LYSMD3
OG0007436	2.89	0.00009	LYSMD4
OG0006021	6.5	0	MALL
OG0007263	7.58	0	MAMDC2
OG0006699	1.28	0	MAPK8IP1
OG0008032	1.06	0.00008	MCF2
OG0006196	1.35	0.00049	MCMDC2
OG0007728	1.69	0.00015	MICAL3
OG0006172	2.12	0.00001	MPP1
OG0006080	3.4	0.03467	MRC2
OG0006277	3.85	0.00001	MRPL46
OG0007071	3.11	0.00582	MRPS9
OG0006109	1.11	0.00016	MSTN
OG0008054	2.17	0	MSTO1
OG0007929	1.47	0	MTF1
OG0008243	3.4	0.00766	MTFR1
OG0007047	1.77	0.00003	MTX2
OG0006883	1.21	0	MYRF
OG0006834	2.12	0	N4BP1
OG0007899	2.66	0	NCF4
OG0006534	3.17	0.00903	NDUFB10
OG0007817	1.01	0	NECTIN1
OG0006597	1.53	0.00059	NFKBIE
OG0007831	4.06	0	NOB1

OG0006928	2.13	0	NPC1
OG0006022	1.78	0	NPHP1
OG0006820	2.17	0	NPHS2
OG0006526	1.1	0	NRDE2
OG0006445	2.01	0	NSMF
OG0007310	3.88	0	NUBP1
OG0007326	2.56	0	NUDCD1
OG0007255	13.1	0	NUFIP1
OG0007022	1.67	0	OMA1
OG0007297	1.36	0.0057	OSGIN2
OG0006518	2.14	0.007	PANX3
OG0008300	1.5	0.00001	PAPPA2
OG0007662	1.55	0.04067	PARP1
OG0006046	1.04	0.00175	PARS2
OG0007725	1.74	0	PCDH1
OG0008065	1.18	0	PCF11
OG0007396	1.44	0.00387	PCNX1
OG0007198	3.47	0.00001	PDCD7
OG0007663	4.21	0	PKDCC
OG0006672	9.89	0	PLAT
OG0008205	4.19	0.0001	PLAU
OG0008333	1.25	0.00001	PLEK
OG0006265	1.11	0.00664	PLEKHB2
OG0007435	6.12	0	PLEKHO2
OG0007412	3.38	0	PM20D1
OG0007837	3.02	0.0001	POF1B
OG0006275	2.22	0	POLG
OG0006889	1.35	0.00411	POP5
OG0007826	2.77	0	PROSER2
OG0007692	3.92	0	PSIP1
OG0006163	1.81	0.00326	PXYLP1
OG0007896	2.8	0.00146	PYROXD1
OG0007634	1.38	0	R3HCC1
OG0008121	1.2	0	RAI1
OG0006241	1.08	0	RARS1
OG0007805	9.47	0.00001	RELL2
OG0006982	1.24	0	RFX6
OG0008199	1.5	0.00037	RGR
OG0006813	1.3	0	RICTOR

OG0007510	1.5	0.00007	RNF146
OG0006329	2.76	0.00052	RPAP1
OG0008147	1.95	0.00016	RRNAD1
OG0006502	1.05	0.00016	RRP7A
OG0008066	1.75	0	RSF1
OG0008149	1.93	0.00745	RTN4R
OG0007502	2.96	0.00074	SERPINH1
OG0008165	1.08	0.00189	SETBP1
OG0007330	2	0	SETX
OG0007984	1.32	0.00025	SGPP2
OG0007865	2.67	0	SGSH
OG0007606	1.78	0	SH3D19
OG0006581	9.14	0	SIK2
OG0006886	3.03	0.02758	SIRT4
OG0008016	1.69	0	SLC16A4
OG0006520	2.11	0.00214	SLC18A3
OG0006574	3.02	0	SLC45A4
OG0007950	2.21	0	SLC46A2
OG0007238	2.53	0	SMO
OG0006227	1.48	0	SMPDL3B
OG0007802	1.96	0.00695	SNX16
OG0008116	2	0	SPATA5
OG0007642	1.76	0	SRCIN1
OG0007260	1.47	0	SSTR5
OG0008251	1.06	0	ST18
OG0006307	2.57	0	SVOPL
OG0008301	2.85	0	SWT1
OG0006993	3.43	0.01893	SYT11
OG0006210	1.09	0	TAB1
OG0006060	3.52	0.01497	TADA2B
OG0006710	2.55	0	TAF4
OG0008074	2.36	0	TBC1D15
OG0006493	1.61	0	TBC1D2B
OG0007521	2.16	0.00353	TBCE
OG0007651	2.74	0.0009	TCF25
OG0006774	1.45	0.00453	TDO2
OG0006017	1.85	0	TDRD9
OG0006322	1.44	0.00011	TECPR1
OG0006803	3.24	0.00012	TGFBI

OG0006024	7.96	0	THY1
OG0006655	1.3	0.00033	TIMM21
OG0007159	1.42	0.00013	TMEM11
OG0006863	1.45	0	TMEM168
OG0007397	1.18	0.00317	TMEM240
OG0007023	2.15	0	TMEM260
OG0007317	2.04	0.00041	TMPRSS3
OG0006246	2.53	0.04086	TNFRSF11B
OG0008022	2.21	0	TNNT2
OG0008311	1.91	0	TNR
OG0007165	2.69	0.00021	TNS4
OG0008213	1.55	0	TOX3
OG0007969	1.04	0.00014	TPTE2
OG0007153	1.35	0.00119	TTC19
OG0007430	3.1	0	UBAP1L
OG0006489	2	0	URB2
OG0007695	2.77	0.00008	USP25
OG0007245	3.17	0.00002	UTP11
OG0007674	1.69	0.00137	UTP25
OG0007328	1.71	0	VIRMA
OG0007456	1.65	0.01769	VLDLR
OG0007074	4.61	0.00001	VSTM2A
OG0008269	7.09	0	VSTM4
OG0006507	2.09	0.00007	WASF1
OG0008316	4.33	0.00002	WBP4
OG0008285	2.48	0.00432	WDR53
OG0006645	1.79	0.00047	WIPI1
OG0006181	1.82	0.00016	WNT16
OG0006505	1.72	0	ZHX2
OG0007089	1.22	0.02872	ZNF277
OG0007431	6.04	0	ZWILCH

Supplementary Table S6. Genome statistics and sources for non-avian reptiles used in this study.

Habitat	Group	Species	Assembly size (Gb)	GC content (%)	Protein	Source	Accession
L	Gastropoda	*Megaustenia siamensis*	2.5	38.24	34882	This study	This study
L	Gastropoda	*Achatina fulica*	1.85	NA	23726	GigaDB	DOI:10.5524/100647
L	Gastropoda	*Arion vulgaris*	1.54	38.46	32518	NCBI	GCA_020796225.1
L	Gastropoda	*Candidula unifasciata*	1.36	40.3	22464	NCBI	GCA_905116865.2
F	Gastropoda	*Biomphalaria glabrata*	0.92	36.1	36675	NCBI	GCA_000457365.1
F	Gastropoda	*Haliotis rufescens*	1.33	40.9	55609	NCBI	GCA_023055435.1
F	Gastropoda	*Pomacea canaliculata*	0.44	40.71	40391	NCBI	GCA_003073045.1
F	Bivalvia	*Dreissena polymorpha*	1.8	35.13	189750	NCBI	GCA_020536995.1
M	Gastropoda	*Aplysia californica*	0.93	41.99	26676	NCBI	GCA_000002075.2
M	Gastropoda	*Chrysomallon squamiferum*	0.46	34.48	28781	GigaDB	DOI:10.5524/100817
M	Gastropoda	*Gigantopelta aegis*	1.29	37.45	25601	GigaDB	DOI:10.5524/100817
M	Gastropoda	*Lottia gigantea*	0.36	36	23822	NCBI	GCA_000327385.1
M	Bivalvia	*Crassostrea gigas*	0.65	33.49	63341	NCBI	GCA_902806645.1
M	Bivalvia	*Mizuhopecten yessoensis*	0.99	33.6	22448	NCBI	GCA_002113885.2

Supplementary Table S7. Repetitive elements in *Megaustenia siamensis*

sequences: 161 total length: 2,593,626,580 bp GC level: 38.24 % bases masked: 1,574,043,348 bp (60.69%)				
Type	SubType	No.of Element	Length (bp)	Proportion of Genome (%)
Retroelements		203642	37581094	1.45
Penelope		6656	816445	0.03
LINEs		98855	15964079	0.62
	L2/CR1/Rex	18763	3793436	0.15
	R1/LOA/Jockey	6106	1076663	0.04
	R2/R4/NeSL	3529	531733	0.02
	RTE/Bov-B	30438	4318208	0.17
	L1/CIN4	22840	3639015	0.14
LTR elements:		102025	21488868	0.83
	BEL/Pao	4955	950902	0.04
	Ty1/Copia	17523	2960241	0.11
	Gypsy/DIRS1	51356	12954469	0.50
	Retroviral	18317	1944650	0.07
DNA transposons		232983	37035194	1.43
	hobo-Activator	45873	8023241	0.31
	Tc1-IS630-Pogo	19504	2768223	0.11
	PiggyBac	1253	162483	0.01
	Tourist/Harbinger	8048	1156016	0.04
Rolling-circles		18023	3973667	0.15
Unclassified:		4762262	1422041331	54.83
Total interspersed repeats			1496657619	57.71
Small RNA		3061	196087	0.01
Satellites		28442	14281899	0.55
Simple repeats		534153	52875196	2.04
Low complexity		56034	6135264	0.24

Supplementary Table S8. 117 unique gene clusters specific in *M.siamensis*

Cluster name	Protein number	Swiss_prot_id	GO_Annotation
cluster101	23	Q9I930	GO:0045088; P:regulation of innate immune response; TAS:UniProtKB
cluster1127	7	P28798	GO:0048488; P:synaptic vesicle endocytosis; ISO:MGI
cluster1128	7	Q5IS39	GO:0019233; P:sensory perception of pain; ISS:UniProtKB
cluster113	21	Q588U8	GO:0005634; C:nucleus; IEA:UniProtKB-SubCell
cluster1131	7	P47820	GO:0042310; P:vasoconstriction; IMP:RGD
cluster1156	7	Q9VZW5	GO:0007204; P:positive regulation of cytosolic calcium ion concentration; IDA:UniProtKB
cluster12103	3	Q9ULY4	GO:0002532; P:production of molecular mediator involved in inflammatory response; IDA:UniProtKB
cluster12108	3	Q7TD08	GO:0003964; F:RNA-directed DNA polymerase activity; IEA:UniProtKB-KW
cluster12109	3	O02751	GO:0005634; C:nucleus; IEA:UniProtKB-SubCell
cluster12111	3	Q01528	GO:0005576; C:extracellular region; IEA:UniProtKB-SubCell
cluster12112	3	Q6GLQ1	GO:0004867; F:serine-type endopeptidase inhibitor activity; IEA:InterPro
cluster12119	3	P08836	GO:0033384; P:geranyl diphosphate biosynthetic process; IEA:UniProtKB-UniPathway
cluster12120	3	Q10751	GO:0003081; P:regulation of systemic arterial blood pressure by renin-angiotensin; IBA:GO_Central
cluster12126	3	Q7JK24	GO:0006487; P:protein N-linked glycosylation; IBA:GO_Central
cluster12129	3	P54120	GO:0009617; P:response to bacterium; IEP:TAIR
cluster12130	3	O02751	GO:0005634; C:nucleus; IEA:UniProtKB-SubCell
cluster12644	3	Q08DY9	GO:0030182; P:neuron differentiation; IBA:GO_Central
cluster12654	3	Q95SX7	GO:0006313; P:transposition, DNA-mediated; IMP:UniProtKB
cluster12656	3	P54120	GO:0009617; P:response to bacterium; IEP:TAIR
cluster12658	3	E9PU17	GO:0006638; P:neutral lipid metabolic process; IEA:Ensembl
cluster12663	3	P62998	GO:0071526; P:semaphorin-plexin signaling pathway; ISS:UniProtKB
cluster12664	3	Q95SX7	GO:0006313; P:transposition, DNA-mediated; IMP:UniProtKB
cluster130	20	Q9H093	GO:0006468; P:protein phosphorylation; IDA:UniProtKB
cluster14320	2	Q9PWA0	GO:0009968; P:negative regulation of signal transduction; IEA:UniProtKB-KW
cluster147	18	O02751	GO:0005634; C:nucleus; IEA:UniProtKB-SubCell
cluster16342	2	Q1E3R8	GO:0000272; P:polysaccharide catabolic process; IEA:UniProtKB-KW
cluster16348	2	P50452	GO:0010951; P:negative regulation of endopeptidase activity; IDA:UniProtKB
cluster16349	2	Q9Z1P7	GO:0051497; P:negative regulation of stress fiber assembly; ISO:MGI
cluster16365	2	Q08821	GO:0006355; P:regulation of transcription, DNA-templated; IEA:InterPro

cluster16366	2	O18992	GO:0005506; F:iron ion binding; IEA:InterPro
cluster16379	2	Q9UMG7	GO:0048251; P:elastic fiber assembly; IMP:UniProtKB
cluster16385	2	Q26636	GO:0007275; P:multicellular organism development; IEA:UniProtKB-KW
cluster16388	2	O02751	GO:0005634; C:nucleus; IEA:UniProtKB-SubCell
cluster16392	2	A4VCL2	GO:0006468; P:protein phosphorylation; IDA:FlyBase
cluster16393	2	Q9I926	GO:0045088; P:regulation of innate immune response; TAS:UniProtKB
cluster16394	2	Q64417	GO:0005506; F:iron ion binding; IEA:InterPro
cluster16395	2	Q8LGG8	GO:0016208; F:AMP binding; IPI:TAIR
cluster16398	2	Q95ZJ1	GO:0018243; P:protein O-linked glycosylation via threonine; IDA:WormBase
cluster16400	2	O75452	GO:0006629; P:lipid metabolic process; TAS:ProtInc
cluster16404	2	O14510	GO:0042060; P:wound healing; TAS:BHF-UCL
cluster166	17	Q49A17	GO:0018243; P:protein O-linked glycosylation via threonine; IDA:UniProtKB
cluster18688	2	P18956	GO:0097264; P:self proteolysis; IDA:EcoCyc
cluster18737	2	Q9VN12	GO:0006814; P:sodium ion transport; ISS:UniProtKB
cluster18739	2	O75452	GO:0006629; P:lipid metabolic process; TAS:ProtInc
cluster18740	2	Q9Z127	GO:0007399; P:nervous system development; IEA:UniProtKB-KW
cluster18742	2	Q03278	GO:0003964; F:RNA-directed DNA polymerase activity; IEA:UniProtKB-KW
cluster18748	2	Q14112	GO:0030198; P:extracellular matrix organization; TAS:Reactome
cluster18749	2	Q4QR71	GO:0051603; P:proteolysis involved in cellular protein catabolic process; IBA:GO_Central
cluster18750	2	Q49A17	GO:0018243; P:protein O-linked glycosylation via threonine; IDA:UniProtKB
cluster18751	2	O60911	GO:0007283; P:spermatogenesis; IEA:Ensembl
cluster18754	2	Q5JWP5	GO:0006465; P:signal peptide processing; IBA:GO_Central
cluster18756	2	P22770	GO:0007271; P:synaptic transmission, cholinergic;
cluster18760	2	Q588U8	GO:0005634; C:nucleus; IEA:UniProtKB-SubCell
cluster18761	2	Q1JPB0	GO:0010951; P:negative regulation of endopeptidase activity; IBA:GO_Central
cluster18763	2	O42603	GO:0007166; P:cell surface receptor signaling pathway; IEA:InterPro
cluster18765	2	O61363	GO:0005344; F:oxygen carrier activity; IEA:UniProtKB-KW
cluster18766	2	Q1LZF1	GO:0003309; P:type B pancreatic cell differentiation; IEA:Ensembl
cluster18768	2	Q9UDJ8	GO:0009617; P:response to bacterium; TAS:ProtInc
cluster18782	2	Q61206	GO:0007283; P:spermatogenesis; IMP:MGI
cluster18784	2	Q99677	GO:0035025; P:positive regulation of Rho protein signal transduction; IBA:GO_Central
cluster18786	2	Q9CX98	GO:0006805; P:xenobiotic metabolic process; IBA:GO_Central
cluster18787	2	P40313	GO:0030163; P:protein catabolic process; IC:GO_Central
cluster18790	2	P12821	GO:0007283; P:spermatogenesis; ISS:BHF-UCL

cluster18795	2	Q9ESG5	GO:0008467; F:[heparan sulfate]-glucosamine 3-sulfotransferase 1 activity; NAS:RGD
cluster18796	2	Q8N0N3	GO:0045087; P:innate immune response; NAS:UniProtKB
cluster18808	2	Q64676	GO:0002175; P:protein localization to paranode region of axon; IMP:BHF-UCL
cluster18809	2	Q8K349	GO:0005525; F:GTP binding; IBA:GO_Central
cluster18812	2	Q8JI28	GO:0007275; P:multicellular organism development; IEA:UniProtKB-KW
cluster18813	2	Q7Z0T3	GO:0005186; F:pheromone activity; IEA:UniProtKB-KW
cluster18816	2	Q7Z449	GO:0006805; P:xenobiotic metabolic process; IBA:GO_Central
cluster18818	2	Q9Y219	GO:0045061; P:thymic T cell selection; IDA:UniProtKB
cluster18819	2	P04323	GO:0015074; P:DNA integration; IEA:InterPro
cluster18824	2	Q05901	GO:0007271; P:synaptic transmission, cholinergic; IBA:GO_Central
cluster18826	2	O08699	GO:0007179; P:transforming growth factor beta receptor signaling pathway; ISS:UniProtKB
cluster18827	2	Q92618	GO:0009409; P:response to cold; IEA:Ensembl
cluster18831	2	P21328	GO:0003964; F:RNA-directed DNA polymerase activity; IEA:UniProtKB-KW
cluster18832	2	O61363	GO:0005344; F:oxygen carrier activity;
cluster18842	2	Q9P2E5	GO:0030206; P:chondroitin sulfate biosynthetic process; TAS:Reactome
cluster18848	2	P12821	GO:0007283; P:spermatogenesis; ISS:BHF-UCL
cluster18849	2	Q8WZB3	GO:0006941; P:striated muscle contraction; TAS:UniProtKB
cluster18857	2	Q8HXQ5	GO:0042908; P:xenobiotic transport; IBA:GO_Central
cluster1947	6	Q10751	GO:0003081; P:regulation of systemic arterial blood pressure by renin-angiotensin; IBA:GO_Central
cluster1950	6	Q60670	GO:0048511; P:rhythmic process; IEA:UniProtKB-KW
cluster20	126	P11369	GO:0006310; P:DNA recombination; IEA:UniProtKB-KW
cluster207	15	Q7Z449	GO:0006805; P:xenobiotic metabolic process; IBA:GO_Central
cluster27	88	O00370	GO:0032197; P:transposition, RNA-mediated; IMP:UniProtKB
cluster3647	5	Q9UMG7	GO:0048251; P:elastic fiber assembly; IMP:UniProtKB
cluster3648	5	Q0C3M1	GO:0006401; P:RNA catabolic process; IEA:UniProtKB-UniRule
cluster3657	5	Q588U8	GO:0005634; C:nucleus; IEA:UniProtKB-SubCell
cluster3660	5	P21328	GO:0003964; F:RNA-directed DNA polymerase activity; IEA:UniProtKB-KW
cluster3662	5	Q95029	GO:0051603; P:proteolysis involved in cellular protein catabolic process; IBA:GO_Central
cluster3736	5	Q7Z449	GO:0006805; P:xenobiotic metabolic process; IBA:GO_Central
cluster3737	5	P79760	GO:0046677; P:response to antibiotic; IMP:AgBase
cluster3738	5	Q95SX7	GO:0006313; P:transposition, DNA-mediated; IMP:UniProtKB
cluster3739	5	A4VCL2	GO:0006468; P:protein phosphorylation; IDA:FlyBase

cluster3743	5	Q9H0Y8	GO:0007169; P:transmembrane receptor protein tyrosine kinase signaling pathway; IMP:UniProtKB
cluster436	10	Q6BD04	GO:0004930; F:G protein-coupled receptor activity; IEA:UniProtKB-KW
cluster4497	4	P77735	GO:0006772; P:thiamine metabolic process; EXP:EcoliWik
cluster48	45	Q0AMI4	GO:0006401; P:RNA catabolic process; IEA:UniProtKB-UniRule
cluster566	9	Q95SX7	GO:0006313; P:transposition, DNA-mediated; IMP:UniProtKB
cluster568	9	Q61847	GO:1901998; P:toxin transport; IMP:MGI
cluster63	35	Q95SX7	GO:0006313; P:transposition, DNA-mediated; IMP:UniProtKB
cluster7356	4	P54120	GO:0009617; P:response to bacterium; IEP:TAIR
cluster7357	4	P25092	GO:0007165; P:signal transduction; IBA:GO_Central
cluster7358	4	Q6P2A1	GO:0010468; P:regulation of gene expression; IBA:GO_Central
cluster7360	4	Q93243	GO:0018996; P:molting cycle, collagen and cuticulin-based cuticle; IMP:WormBase
cluster7364	4	P30740	GO:0043312; P:neutrophil degranulation; TAS:Reactome
cluster7368	4	Q8MRC9	GO:0006493; P:protein O-linked glycosylation; ISS:FlyBase
cluster7369	4	Q05909	GO:1901998; P:toxin transport; IMP:MGI
cluster7370	4	Q7Z449	GO:0006805; P:xenobiotic metabolic process; IBA:GO_Central
cluster7602	4	Q4V8D1	GO:0006805; P:xenobiotic metabolic process; IBA:GO_Central
cluster7606	4	Q964E1	GO:0005524; F:ATP binding; IEA:UniProtKB-KW
cluster774	8	Q9VZW5	GO:0007204; P:positive regulation of cytosolic calcium ion concentration; IDA:UniProtKB
cluster775	8	Q5JXM2	GO:0008168; F:methyltransferase activity; IEA:UniProtKB-KW
cluster776	8	Q588U8	GO:0005634; C:nucleus; IEA:UniProtKB-SubCell
cluster89	26	Q69ZU8	GO:2000096; P:positive regulation of Wnt signaling pathway, planar cell polarity pathway; IDA:MGI
cluster92	25	Q95SX7	GO:0006313; P:transposition, DNA-mediated; IMP:UniProtKB

Supplementary Table S9. Function enrichment of specific gene families in *M.siamensis*

GO_ID	Name	Namespace	Count	p-value
GO:0005634	nucleus	cellular_component	8	4.16E-07
GO:0006313	transposition, DNA-mediated	biological_process	6	1.78E-08
GO:0006805	xenobiotic metabolic process	biological_process	6	2.96E-05
GO:0003964	RNA-directed DNA polymerase activity	molecular_function	4	5.56E-06
GO:0009617	response to bacterium	biological_process	4	0.000368
GO:0018243	protein O-linked glycosylation via threonine	biological_process	3	0.000632
GO:0006468	protein phosphorylation	biological_process	3	0.006632
GO:0048251	elastic fiber assembly	biological_process	2	0.000551
GO:0010951	negative regulation of endopeptidase activity	biological_process	2	0.000909
GO:0045088	regulation of innate immune response	biological_process	2	0.001351
GO:0003081	regulation of systemic arterial blood pressure by renin-angiotensin	biological_process	2	0.001351
GO:0005506	iron ion binding	molecular_function	2	0.001873
GO:0006629	lipid metabolic process	biological_process	2	0.003899
GO:0006401	RNA catabolic process	biological_process	2	0.00472
GO:1901998	toxin transport	biological_process	2	0.00472
GO:0005344	oxygen transporter activity	molecular_function	2	0.00561

Supplementary Table S10. 17 orthologous groups with signatures of positive selection in *M.siamensis*

Orthogroup	dn/ds	p-value	Gene
OG0011944	8	6.68E-08	AK6
OG0012045	4.6	0.0000114	RPL6
OG0012047	10	0	FTSJ3
OG0012086	6.1	3.52E-12	FEN1
OG0012116	4.9	0	FBXW7
OG0012134	9.8	4.47E-13	SGPL1
OG0012140	5.8	0	SRP68
OG0012171	2.1	0.00000282	CHMP3
OG0012254	10	0.00000125	PSMG2
OG0012257	13	1.19E-12	DERL1
OG0012313	19	4.66E-15	IRAK1BP1
OG0012328	9.5	0.000000416	-
OG0012417	5.8	0.000000358	PSMB7
OG0012419	1.2	0.00083	COPS5
OG0012425	3.9	0.00000208	PSMD1
OG0012429	6.6	0	SMC2
OG0012431	7.2	0	NOL10

VITA

NAME	Wanna Chetruengchai
DATE OF BIRTH	28 June 1993
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	2015-2018 M.Sc in Bioinformatics and Systems Biology Program, School of Bioresource and Technology and School of Information Technology, King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, Thailand. 2011-2015 B.Sc. Industrial Microbiology (2nd class honor), King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok, Thailand.
HOME ADDRESS	Motto Kanchanapisek-Rama 2, Kanchanaphisek Road, Bangbon, Bangbon, Bangkok, 10150
PUBLICATION	Boonsimma P, Ittiwut C, Kamolvisit W, Ittiwut R, Chetruengchai W, Phokaew C, Srichonthong C, Poonmaksatit S, Desudchit T, Suphapeetiporn K, Shotelersuk V. Exome sequencing as first-tier genetic testing in infantile-onset pharmacoresistant epilepsy: diagnostic yield and treatment impact. European Journal of Human Genetics. 2022 Oct 5:1-9. Sinthuwiwat T, Buranapraditkun S, Kamolvisit W, Tongkobpetch S, Chetruengchai W, Srichomthong C, Assawapitaksakul A, Phokaew C, Kueanjinda P, Palaga T, Boonpiyathad T. A LILRB1 variant with a decreased ability to phosphorylate SHP-1 leads to autoimmune diseases. Scientific reports. 2022 Sep 14;12(1):1-1. Wankaew N, Chariyavilaskul P, Chamnanphon M, Assawapitaksakul A, Chetruengchai W, Pongpanich M, Shotelersuk V. Genotypic and phenotypic landscapes of 51 pharmacogenes derived from whole-genome sequencing in a Thai population. PloS one. 2022 Feb 17;17(2):e0263621. Chetruengchai W, Shotelersuk V. Actionable secondary findings in the 73 ACMG-recommended genes in 1559 Thai exomes. Journal of Human Genetics. 2022 Mar;67(3):137-42.