WHOLE-GENOME SEQUENCING, *DE NOVO* ASSEMBLY AND COMPARATIVE ANALYSES OF *MANIS JAVANICA* AND *MANIS PENTADACTYLA*

RANJEEV HARI

FACULTY OF DENTISTRY

UNIVERSITY OF MALAYA

KUALA LUMPUR

2017

WHOLE-GENOME SEQUENCING, *DE NOVO* ASSEMBLY AND COMPARATIVE ANALYSES OF *MANIS JAVANICA* AND *MANIS PENTADACTYLA*

RANJEEV HARI

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF DENTISTRY

UNIVERSITY OF MALAYA

KUALA LUMPUR

2017

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: RANJEEV A/L HARI

Registration/Matric No: DHA120020

Name of Degree: Doctor of Philosophy (PhD)

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Whole-genome sequencing, de novo assembly and analyses of *Manis javanica*

Field of Study: PhD (Bioinformatics)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every right in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:

Designation:

ABSTRACT

Pangolins are unique mammals with scales over most of their body and they have no teeth, have poor vision, and an acute olfactory system. Pangolins comprise the only placental order (Pholidota) whose genomes have not been completely sequenced. The living Pholidotans are comprised of four Asian (*Manis javanica*, *Manis pentadatyla*, *Manis culioensis*, *Manis. crassicaudata*) and four African (*Manis temmincki*, *Manis gigantae*, *Manis tricuspis*, *Manis tetradactlya*) species. In this study, I sequenced and assembled the first reference genome of the Malayan pangolin (*M. javanica*), which is a critically endangered species from Malaysia and compared it with the closely related Chinese pangolin (*M. pentadactyla*) from Taiwan. Using the MAKER annotation pipeline, I identified 23,446 and 20,298 protein-coding genes in the *M. javanica* and *M. pentadactyla* genomes respectively. Strikingly, pangolins have a non-canonical repeat layout compared to closely related Carnivorans. Pangolins lack tRNA-SINES and have short mean intron length (likely due to the loss of a tRNA-SINE family) compared to other mammals, which may be associated to an inherent metabolic adaptation in the pangolins. Furthermore, ancient population history modelling showed the dwindling population of endangered pangolins, providing important insights for future conservation efforts. The results showed that long-term population decline predated recent declines which highlights the urgency to stunt such trends by introducing tighter regulations and accelerating conservation efforts of these endangered pangolins. I also found the presence of *Burkholderia fungorum* in the supposed sterile pangolin tissues of cerebrum, cerebellum, lung and blood, suggesting that this bacterium has capability to colonise pangolins. In addition to the nuclear genomes of pangolins, I also assembled the associated reference mitogenomes of *M. javanica and M. pentadactyla*. Phylogenetic trees based on the newly assembled mitogenomes confirmed that the pangolin that were sequenced in this study were indeed *M. javanica*. Furthermore, I have also identified and annotated 7,422 expressed long non-coding RNA (lncRNA) in pangolins. Overall, the pangolin whole-genome assemblies, mitogenomes, list of genes and lncRNAs will serve as important resources to the research community in the future in areas such as conservation efforts, molecular biology, genetics and evolutionary biology.

ABSTRAK

Tenggiling merupakan mamalia unik bersisik keratin yang meliputi kebanyakan badannya, tidak bergigi, mempunyai penglihatan yang lemah, dan mempunyai sistem olfaktori yang tajam. Tenggiling adalah dari ordo plasenta (Pholidota) yang genomnya belum sepenuhnya dijujuk. Anggota Pholidota ini terdiri daripada empat spesies Asia (*Manis javanica*, *Manis pentadatyla*, *Manis culioensis*, *Manis. crassicaudata*) dan empat spesies Afrika (*Manis temmincki*, *Manis gigantae*, *Manis tricuspis*, *Manis tetradactlya*). Dalam kajian ini, saya telah menjujuk dan menyusun genom rujukan pertama tenggiling Malaya (*M. javanica*) dari Malaysia, yang dikhuatiri terancam dan telah dibandingkan dengan spesies berkait terdekat iaitu tenggiling Cina (*M. pentadactyla*) dari Taiwan. Melalui perisian MAKER, saya dapat mengenalpasti 23,446 dan 20,298 gen mengekod protein di genom *M. javanica* dan *M. pentadactyla* masing-masing. Yang menarik, tenggiling didapati mempunyai turutan berulang nukleotida yang agak berbeza dengan familia Carnivora yang terdekat dengan Pholidota. Tenggiling yang didapati kekurangan turutan berulang tRNA-SINES selain min intron yang pendek (mungkin disebabkan oleh ketiadaan famili tRNA-SINE sendiri) berbanding mamalia lain boleh dikaitkan dengan kewujudan suatu penyesuaian metabolik yang unik kepada tenggiling. Tambahan pula, pemodelan populasi purba (PSMC) tenggiling yang menunjukkan populasi tenggiling yang semakin menjunam dapat memberikan petanda keafiatan populasi tersebut. Hasil kajian menunjukkan bahawa penurunan populasi purba mendadak telah mendahului penurunan baru-baru ini sekaligus memberi petanda yang penting untuk membantutkan trend tersebut melalui undang-undang yang lebih ketat serta meningkatkan usaha konservasi tenggiling yang terancam. Saya juga mendapati kehadiran *Burkholderia fungorum* dalam tisu tenggiling yang sepatutnya steril seperti pada serebrum, serebelum, paru-paru dan darah, yang sebaliknya menunjukkan bahawa bakteria ini mempunyai keupayaan untuk menjakiti tenggiling. Selain genom nuklear tenggiling, saya juga menyusun kembali genom mitokondria dengan menggunakan genom rujukan *M. javanica* dan *M. pentadactyla*. Cabangan filogenetik berdasarkan genom mitokondria yang baru disusun mengesahkan spesies dalam kajian ini sememangnya *M. javanica*. Selain itu, saya juga telah mengenal pasti serta menganotasi 7422 transkrip nyah-pengekodan RNA panjang (lncRNA) tenggiling. Secara keseluruhan, penjujukan genom tenggiling, genom mitokondria, dan senarai gen serta lncRNA dalam kajian ini merupakan sumber penting untuk penyelidik pada masa akan datang dalam bidang seperti usaha-usaha konservasi, biologi molekul, genetik dan biologi evolusi.

ACKNOWLEDGEMENT

This research work would not have been possible without the support of many people. I am extremely indebted to my supervisor Dr Lawrence Choo Siew Woh, whose encouragement, guidance, support and priceless assistance from the start to the concluding stage enabled me to widen my knowledge in this field and made me breeze through the trials faced during my period of work. I have my deepest appreciations also to Prof. Ian Patterson in which without his guidance and assistance this study would not have been successful. I distinctively wish to convey my appreciation to Hazhir Hajian and Tan Shi Yang for their assistance and help in server handlings and, Aini Yasmin, Winnie Wong Guat Jah, Tan Tze King, Tan Ka Yun that helped in my early research undertakings. My gratitude also goes to our collaborators and veterinarians at PERHILITAN who helped us so much with sampling. I am also especially thankful to the International Pangolin Research Consortium (IPaRC) members both local and abroad. A big round of thanks goes to all colleagues of Genome Informatics Research Group (GIRG), Faculty of Dentistry lab members, administration staffs of High Impact Research for their invaluable help, expertise and excellence besides being patient at all times in my dealings. Special thanks also go to my dear wife Suhanya Parthasarathy whose experiential advise and continuous encouragement throughout the research period reassured my composure in achieving my dream. This research was only probable due to the monetary support provided by University of Malaya research grant. I would also like to extend my appreciation to the Ministry of Education for providing me the MyPhD scheme which relieved me the financial burden during the period of research. Special gratitude is obliged to my late father, Hari and mother, Sreedevi who have moulded, encouraged and supported me over the years in all way irrespective of the field of study and distance I came to pursue my dream in this area of bioinformatics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

BGI	Beijing Genome Institute
cDNA	complementary DNA
CNV	copy number variation
CR	control region
CYTB	Cytochrome B
DBG	de-Bruijn Graph
DHU	dihydrouridine
DNA	deoxy-nucleic acid
DWNPM	Department of Wildlife and National Parks Malaysia
emPCR	emulsion PCR
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
FRC	Feature Response Curve
HMW	high molecular weight
IACUC	International Animal Care and Use Committee
KYA	thousand years ago
LD	Linkage disequilibrium
lincRNA	long intergenic RNA
lncRNA	long non-coding RNA
mRNA	messenger RNA
MYA	millions of years ago
NCBI	National Centre For Biotechnology Information
Ne	Effective population size
NGS	next-generation sequencing
OLC	overlap-layout-consensus
ORF	open reading frame

PCG	protein-coding genes
PCR	Polymerase chain reaction
PSMC	Pairwise Sequential Markovian Coalescent
RABT	Reference Annotation Based Transcript
RCSU	relative synonymous codon usage
RNA	ribo-nucleic acid
SD	segmental duplication
SNP	single nucleotide polymorphism
SRA	Sequence Read Archive
TCM	Traditional Chinese Medicine
TE	transposable elements
tRNA	transfer RNA
UV	ultraviolet
UV-Vis	UV visible
WGAC	whole genome alignment comparison
WGS	whole genome sequencing

CHAPTER 1: INTRODUCTION

1.1 Background

Pangolins (Order Pholidota; Family Manidae) are animals which have fascinated man for a very long time. The intriguing keratinous armour that surrounds the body and its behaviour to curl up to a ball when provoked are the profound characteristics that are commonly recognized (Jentink, 1903). Pangolins were once grouped with sloths, giant anteater and armadillos under Edentata based loosely on anatomy and general aspects of biology and behaviour (Cuvier, 1833). However, later studies on systematics using molecular DNA evidence supported the grouping into a separate order of Pholidota with the closest living relatives being the diverse order of Carnivora (Murphy et al., 2001).

Largely, the focus on pangolin research has been on anatomical, breeding and behavioural studies (Nisa, 2005; Ofusori et al., 2007; Prapong et al., 2009; Rachmawati, 2011; Ruhyana, 2007; Mohapatra & Panda, 2014; V. T. Nguyen et al., 2010; C. W. Yang et al., 2007). While other studies have focussed on pangolin populations (S. Wu, Liu, Zhang, & Ma, 2004) and natural parasites (Mohapatra, Panda, Nair, & Acharjyo, 2015). Notably, complex genetic studies are still lacking.

One study characterized the sex determination gene, *Sry* (Yu et al., 2011) and several others have examined the mitochondrial DNA of pangolins (Du Toit et al., 2014; Hassanin, Hugot, & van Vuuren, 2015; Qin, Dou, Guan, Qin, & She, 2012; Semiadi & others, 2013). However, genetic information such as the genome size and genetic markers derived from microsatellites or single nucleotide polymorphisms (SNPs) were critically absent.

The genome of pangolin thus allows a rare opportunity and functions as an important resource for comparative genomics studies. By sequencing a single genome, it offers the possibility of identifying thousands of genes and regulatory sequences, enables the surveying of genome architecture, and delineation of complex functional pathways of genetic basis. In short, it would allow a better understanding of mammalian genome evolution and could further support and provide evidence relating to genetic functions determined in other whole-genome sequencing projects, particularly the human genome.

1.2 Brief biological rationale for sequencing the pangolin genome

Pangolins exhibit a number of unique traits among mammals that lay the foundation for interesting comparative genomics studies. First, pangolins are adapted well for a myrmecophagous lifestyle with diets consisting exclusively of ants and termites. The dietary adaptations can be observed in the form of robust forelimbs, enlarged claws, and a long, sticky tongue for digging out nests and capturing insects (Kingdon, 1971; Skinner & Smithers, 1990). Secondly, the pangolins possess hardened keratinous scale all over their body as armour for protecting them against predators. They also have been noted to burrow underground to keep warm in cold nights. Thirdly, pangolins lack teeth and instead possess a gizzard, a modification in their stomach cavity with keratinized spines. This structure contains small stones and sand to help mechanically digest their prey (Davit-Béal, Tucker, & Sire, 2009; Krause & Leeson, 1974). As outlined, the pangolins have a rich set of evolutionary adaptations that are unique and represent a potential area of study. The genome sequence could be used to study the gene(s) responsible for the unique traits such as the keratinous scales, appearance of gizzard-like structures and thermo-regulation in pangolins. More generally, generating a catalogue of annotated genes from whole-genome sequencing would also serve as a starting point for advanced genetic analysis.

As per phylogenetics, pangolins were earlier thought to belong to the Edentata group that describes mammals that exhibit toothlessness (Cuvier, 1833). However, many studies have identified that pangolins form their own order Pholidota with their closest relative being the Carnivora order (Arnason et al., 2002; Delgado, Vidal, Veron, & Sire, 2008; Yu et al., 2011), yet their divergence remains unknown. In molecular phylogenetic terms, estimation of divergence time is important for fitting the value into other phylogenomic models such as the Pairwise Sequential Markovian Coalescent (PSMC) model. Apart from that, the associated mitochondrial DNA (mitogenome) of the pangolin can also be used for phylogenetic analysis to confirm the species being investigated.

Ecologically, pangolins are an important species as they are one of the major consumers of ants and termites. Unfortunately, pangolins are in the International Union for Conservation of Nature (IUCN) Red List of critically endangered animals as it has been widely captured illegally for the exotic meat trade and ethno-medical usage in Traditional Chinese Medicine (TCM) (Xu, Zong, & Li, 2009; L. Zhang, Hua, & Sun, 2008). There have been strong legal efforts to curtail the problem but efforts to breed pangolin in captivity have often failed as they are found to die prematurely (C. W. Yang et al., 2007).

1.3 Objectives

The objectives of this study are:

1. To sequence and assemble the nuclear and mitochondrial genome (mitogenome) of *M. javanica* ;
2. To annotate the various structural and genetic features of the assembled nuclear genome;

3. To assemble and analyse the mitogenomes of *M. javanica* and *M. pentadactyla* pangolins;
4. To investigate the evolutionary biology of *M. javanica* and *M. pentadactyla* in regards to the divergence time, phylogeny across *Manis spp*. and explore models to better understand the evolution of the pangolins; and
5. To identify and analyse long non-coding RNA (lncRNA) of pangolins by integrating genome data and its associated expression data from RNA-seq

CHAPTER 2: LITERATURE REVIEW

2.1 What are pangolins?

The pangolins, commonly known as the scaly anteaters, are a unique species. It earned its name from the Javan word, *Pangoeling* because of its peculiar behaviour of rolling into a ball when encountering threat (Cuvier, 1833). The term was thought to be first coined in 1734 by Dutch sailors.

Pangolins are an unusual group of mammals characterized by an outer covering of thorny, overlapping keratinous scales and the absence of teeth (T. J. Gaudin, 2004; Pocock, 1924; Skinner & Smithers, 1990). They are adapted to a myrmecophagous lifestyle, specialized for feeding on ants and termites. In association with this dietary adaptation, they exhibit robust forelimbs, enlarged claws, and a long, sticky tongue for digging out nests and capturing insects (Kingdon, 1971; Skinner & Smithers, 1990). Eight living species of pangolins are known from Asia and Africa, all assigned to the family Manidae. Four species are known from Africa: the four-toed arboreal pangolin *M. tetradactyla*, the arboreal pangolin *M. tricuspis*, the giant pangolin *M. gigantea* (all co-existing in western and central Africa), and the common pangolin *M. temminckii* (distributed in southern and eastern Africa) (Emry, Skinner, & others, 1970; McKenna, Bell, & Simpson, 1997; Skinner & Smithers, 1990). The four living Asian species are *M. crassicaudata*, *M. pentadactyla*, *M. javanica* and *M. culionensis* (Gaubert & Antunes, 2005; Timothy J. Gaudin & Wible, 1999; McKenna et al., 1997).

2.2 Ancient records of pangolins

The earliest records of the pangolins were found in the Shennong's Herbal Atlas (Shennong Bencao Jing) written during the reign of the Qin and Han dynasties (221 BC

to 220 AD) which shows the use of pangolins in traditional Chinese medicine (Shen, 1996). Later in 1767, records were found in Western literature as the pangolin were recognized and documented as quadrupeds and sometimes known as the Javan Devil of the Javan islands (Collini, 1767).

The first positive fossil remains of the pangolin in South Africa were documented in 1976 by Hendey and was assigned as *M. gigantae* that was dated five million years ago (Botha & Gaudin, 2007). The oldest extinct ancestor of pangolin found thus far was from the Eocene (50 MYA) of Germany (Storch, 1978; Storch & Martin, 1994).

2.2.1 Modern pangolin species

As outlined in the previous section, there are eight living pangolin species of which four are the African species (*M. gigantae*, *M. temmincki*, *M. tricuspis* and *M. tetradactyla*) while the other four are the Asian species (*M. javanica*, *M. pentadactyla*, *M. craussicadata* and *M. culionensis*) (Table 2.1; Figure 2.1).

Table 2.1 Asian pangolin species and their distribution by country.

Species	**Range**
Indian Pangolin (*M.* crassicaudata)	India, Pakistan, Sri Lanka, possibly Bangladesh
Malayan Pangolin (*M. javanica*)	Brunei, Indonesia, Lao PDR, Malaysia, Myanmar, Thailand, and Vietnam
Chinese Pangolin (*M. pentadactyla*)	China (including Taiwan), India, Lao PDR, Myanmar, Nepal, Thailand and Vietnam
Palawan Pangolin (*M. culioensis*)	Philippines

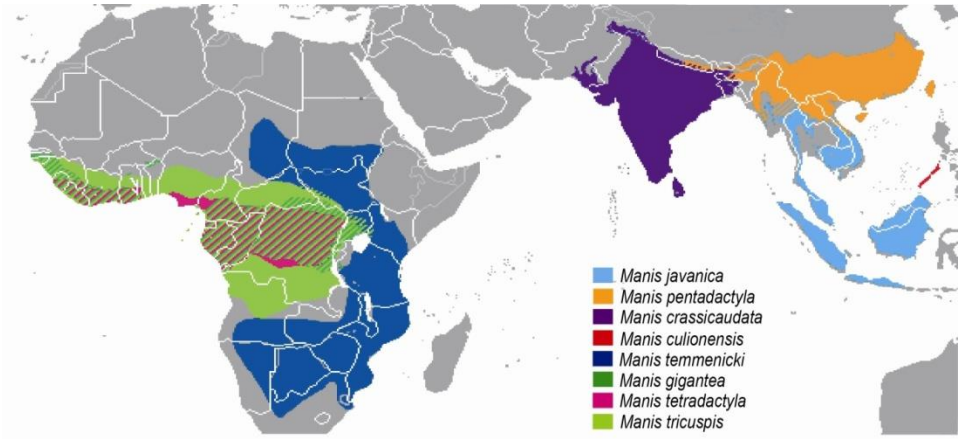

Figure 2.1 Distribution map of species of pangolin.
Geographical distribution maps of the pangolins. There are eight known pangolin species represented by different colours in the map. The hash mark on the map means overlap of geographic areas where more than one species inhabits the same range. Adapted from **(Pemberton, 2011)**.

Morphological identification of pangolin is crucial especially in research and tagging of captured pangolins by wildlife personnel for forensics, records and reporting. Recent molecular analysis showed that species misidentification took place in both the Asian and African species (Hassanin et al., 2015). While pangolins are not established as a cryptic species, their identification may be sometimes difficult due to inexperience in differentiating highly similar morphologies between the pangolin species (Figure 2.2).

Figure 2.2 External morphologies of various pangolin species.
Following is corresponding species to the label: 1. Cape Pangolin or Temminck's Ground Pangolin 2. Chinese Pangolin 3. Sunda Pangolin or Malayan Pangolin 4. Philippine Pangolin 5. Giant Ground Pangolin 6. Indian or Thick-tailed Pangolin 7. Tree Pangolin or African White-bellied Pangolin 8. Long-tailed or Black-bellied Pangolin. (*Handbook of the Mammals of the World, Vol. 2*, 2011).

There are some features common to all pangolins such as their large claws to dig into nests of colonial insects, the absence of teeth, keratinous scales as body armour, tapered snout and its extrusible tongue. The Asian pangolins are sometimes distinguished from African pangolins by the presence of bristles in between the scales (Gray, 1865).

2.2.2 Morphological identification and differences in Asian pangolins

Morphological identification of Asian pangolins is challenging and requires expert knowledge to discriminate minute morphological differences between them (Gaubert & Antunes, 2005). Compared to *M. javanica*, *M. pentadactyla* are known to have fewer than 21 scales along the edge of their tails and the claws on their rear feet is smaller than on the fore feet (V. T. Nguyen et al., 2010). In fact, the claw length ratio should be more than 0.5 in *M. javanica*. The protruding rim of the ear in *M. pentadactyla* is more than 10 mm while in *M. javanica* is lesser (S. Wu, Liu, Zhang, Ou, & Chen, 2003).

2.3 Threats to pangolins

Prior to 2010, the Asian pangolin species were categorised as vulnerable, however, later the status has changed to "Critically Endangered" for Malayan and Chinese pangolin while Endangered for Indian and Philippine pangolin. According to a study by Wu and colleagues (S. Wu et al., 2004), in China the pangolin population has plummeted by up to 94% since 1960. There are several factors that are thought to have caused pangolins to be threatened such as symbolic gifting practices, dubious ethnomedical practices, ecological effects and sensitive pangolin physiology that affects its survival but mainly it is attributed due to rampant illegal trade. Wild life trafficking is estimated to be worth $8 billion annually second only to illegal drug trading.

2.4 Genomics in genetic conservation

Primarily the applications of genomics to conservation biology can be broadly categorised into comparative genomics and adaptation genomics (Angeloni, Wagemaker, Vergeer, & Ouborg, 2012; Kohn, Murphy, Ostrander, & Wayne, 2006; Koonin & Galperin, 2003; Stapley et al., 2010). As in the comparative genomics landscape, the increasing number of species with available genomes has enabled comparative studies of genome structure, function, and evolution. Indeed, increasingly new genome assemblies of various species are released annually, for example the international Genome 10KProject that aims to sequence 10,000 vertebrate genomes (Haussler et al., 2009). The next-generation sequencing (NGS) revolution allows new integrative approaches combining comparative and population genomics. Using population genetics models, a single diploid genome can inform about ancient demographic events and assign population ancestry (H. Li & Durbin, 2011). Previously, primate comparative genomics with human allowed the mapping of CNVs and segmental duplications (Prüfer et al., 2012).

In relation to evolutionary adaptation, the explosion of NGS technologies have promoted the use of comparative genomics approaches to study the adaptation and selection in far more detail than previously possible, which has potential implications for conservation (Kohn et al., 2006). As examples, NGS data have been used to identify positive selection (Yi et al., 2010), identify and analyse microsatellites markers linked to immune related genes (Mäkinen, Vasemägi, McGinnity, Cross, & Primmer, 2015), attempt to associate genotype and phenotype by coupling transcriptomic approach (Ekblom, Farrell, Lank, & Burke, 2012), and evolution of genes controlling sexual dimorphism (Moghadam, Pointer, Wright, Berlin, & Mank, 2012). Examples of studies, animal sequenced and other applications of conservation genomics are summarized in Table 2.2.

Table 2.2 Whole genome sequencing applications on conservation of threatened species

Analyses	Example of Animal	Conservation impact	Reference
Selection	*Peromycus,* Oscillated lizards	Adaptation to captivity	(Linnen, Kingsley, Jensen, & Hoekstra, 2009; Nunes, Beaumont, Butlin, & Paulo, 2010)
Genetic variation	Bison, House sparrow	Insights into population history, inbreeding, demography, gene flow, effective population size, and disease susceptibility	(Pertoldi et al., 2009; Wetzel, Stewart, & Westneat, 2011)133, 134
Population genomics	Wolf, Chimpanzee	Determine demographic history of population	(Bowden et al., 2012; vonHoldt et al., 2011)
Inbreeding/Outbreeding depression	Wolf, Bighorn sheep, Florida panther	Genetic rescue and translocation of individuals into the wild	(Hagenblad, Olsson, Parker, Ostrander, & Ellegren, 2009; W. E. Johnson et al., 2010; Stapley et al., 2010)
Hybridisation	Bison, Carrion crow	Assess levels of introgression for population management	(Halbert, Ward, Schnabel, Taylor, & Derr, 2005; Wolf et al., 2010)
Phylogenetics	Ruminants, Primates	Identify species, conservation units, establish phylogenetic relationship among taxa of conservation	(Decker et al., 2009; Perelman et al., 2011)
Phylogeography	Pitcher plant mosquitoes, Birds	Identify phylogeographic patterns and define conservation units	(Emerson et al., 2010; McCormack et al., 2013)
Comparative genomics	Gorilla	Identify the genetic basis of traits, selective pressures in genetic regions	(Scally et al., 2012; Ventura et al., 2011)

*From Steiner (2013)

2.5 The current status of pangolin research

Pangolin research is being actively pursued in a limited number of disciplines. Behavioural research was undertaken in Singapore and histomorphological studies have been performed in Africa, Thailand and Indonesia (Nisa, 2005; Ofusori et al., 2007; Prapong et al., 2009; Rachmawati, 2011; Ruhyana, 2007). In Vietnam, Taiwan, and India extensive documentation and research have been undertaken on captive breeding (Mohapatra & Panda, 2014; V. T. Nguyen et al., 2010; C. W. Yang et al., 2007). A recent review of captive breeding describes the short-lived success of a few cases in producing offspring from these programs but there was no success of producing the next filial generation (Hua et al., 2015). The researchers often cite diet, environmental stress, and predisposition to disease such as gastric ulcer and pneumonia due to poor immune as main causes for the failure.

Research into pangolin conservation genetics is still in its infancy. Currently only a few fully sequenced mitochondrial DNA (mitogenome) sequences of various pangolin species is available. Moreover, a whole-genome map of pangolins is not yet available and, therefore, an opportunity exists to sequence the whole-genome of pangolins that can be used to further understand the biology, genetics, evolution, phylogeny, genetic variations and genome organisation of the pangolins. Such analyses would be novel and add considerably to the body of knowledge and may further enhance the conservation of these endangered species.

2.6 Biological rationale for sequencing the pangolin genome

2.6.1 A unique mammal for comparative studies

The pangolins exhibit a number of unique traits among mammals that lay the foundation for interesting comparative genomic studies. First, pangolins are adapted well for a myrmecophagous lifestyle with diets consisting exclusively of ants and termites. In association with this dietary adaptation, they exhibit robust forelimbs, enlarged claws, and a long, sticky tongue for digging out nests and capturing insects (Kingdon, 1971; Skinner & Smithers, 1990). Second, they scurry away quickly to escape predators, however, under imminent threat they rely on the hardened keratinous scale all over their body as armour by rolling up into a ball. Third they are nocturnal and fourth they are edentulous animals lacking any real or pseudo teeth (Kingdon, 1971; Skinner & Smithers, 1990; Strandberg, 1918).

Due to the lack of teeth, pangolins possess modifications in their stomach cavity that are gizzard-like pyloric region with keratinized spines and contain small stones and sand to help mechanically digest their prey (Davit-Béal et al., 2009; Krause & Leeson, 1974). A previous attempt to PCR amplify one of the genes responsible for tooth development (*DPP* gene) was unsuccessful (McKnight & Fisher, 2009). The study could not identify the reason of the failure as they could not rule out absence of gene, extensive pseudogenization or primer flaw as cause.

A whole-genome sequence could be used to study the gene responsible for the unique traits and causes of mutations affecting such tooth development genes could be identified. More generally, generating a catalogue of annotated genes from whole-genome sequencing would also serve as a starting material for advanced genetic analysis.

2.6.2 Pholidotan and evolutionary relationship

Initially, pangolins were thought to belong to the Edentata group, mainly formed to describe mammals that exhibited toothlessness (Cuvier, 1833). Later, many studies using molecular biology and bioinformatics techniques targeting various genes reveal that pangolins form their own order Pholidota with their closest relative being the Carnivora order (Arnason et al., 2002; Delgado et al., 2008; Yu et al., 2011). Nevertheless, the time of divergence between this two orders are unknown. While fossil evidence suggest the appearance of the pangolin ancestor since 50 MYA in Europe (Timothy J. Gaudin & Wible, 1999; Gebo & Rasmussen, 1985), an accurate molecular systematic method has yet been performed. Estimation of divergence time is important as it can be used to compare the evolution of certain genes and used in phylogenomic models. Annotations from whole-genome sequencing data can be used to find conserved 4-fold degenerate sites in orthologous genes across many species to construct a tree. Coupled with fossil-calibrated divergence times of species, divergence time can be estimated.

The associated mitochondrial DNA (mitogenome) of the pangolin also can be used to confirm the identity of the species being investigated. Besides that, it can add to the pool of genetic resources already available in public databases which has applications in forensic and legal use where illegal trafficking is concerned.

2.6.3 Unique features of the pangolin genomes

The *M. pentadactyla* genome is estimated to be 3.48 pg according to one study and only slightly smaller than the 3.5 pg size of human genome (Yu et al., 2012). While the sample in that study had a diploid number of 40, there are other reports that range from 36 to 42 (Che et al., 2007; Nie, Wang, Su, Wang, & Yang, 2009; Yu et al., 2011). On the other hand, genome size of *M. javanica* is unknown but the diploid number was found to be 38

(Nie et al., 2009). The karyotype of *M. javanica* and *M. pentadactyla* when compared, presented with seven Robertsonian rearrangements that differ between the species. Besides karyotype information, the content of transposable elements, number of SNPs and heterozygosity is yet to be explored.

2.6.4 Potential medicinal and ecological value

Ecologically, pangolins are an important species as they are one of the major consumers of ants and termites. Thus, pangolin population size affects the health of the tropical ecology extensively. Presently, this animal is in the IUCN Red List of endangered animal as it has been widely captured illegally for the exotic meat trade and ethno-medical usage in TCM (Xu et al., 2009; L. Zhang et al., 2008). Indirectly, the enforcement and raids to curtail the trade has an economic cost for countries involved. Efforts to breed pangolin in captivity have often failed as they are found to die prematurely (C. W. Yang et al., 2007).

While ethno-medical claims are yet to be proven, the availability of the genome could accelerate potential discoveries for the pharmaceutical industry. Sequencing the pangolin genome will identify genes and regulatory pathways that are highly conserved which can be exploited for novel biosimilars and therapeutics. Similar expression profiles of either coding or non-coding RNA transcripts across pangolins and human or murine models, can inform ancient and essential functions of interest to pharmaceuticals as important genes that inform drug metabolism besides as a possible source of antibiotics or antivirals.

2.7 Overview of Next Generation Sequencing (NGS) technologies

The era of NGS began following the first reports of this innovative technique in 2005 (Margulies et al., 2005; Shendure & Ji, 2008). The high-throughput NGS technology which uses parallel amplification and sequencing yields shorter read lengths and average raw error rates of 1-1.5% (Shendure & Ji, 2008) when compared to conventional Sanger sequencing protocols which makes up to 1000bp of 99.999% per base accuracies, Nevertheless, bases that have inaccuracy less than 0.1% can be carefully chosen algorithmically.

There are many advantages of NGS, compared to the conventional Sanger sequencing, such as: (i) NGS permits a considerably higher degree of parallelism than Sanger's conventional capillary-based sequencing. (ii) removal of bottlenecks and limitations in previous methods (*E. coli* transformation and colony picking) by *in vitro* construction of a sequencing library, followed by in vitro clonal amplification further enforces parallelism in the workflow (iii) NGS system usually uses reagents that operate on immobilized target on arrays in microliter quantities, which costs are gained back over the full set of sequencing features on the array because the volume per feature are in the range of picoliters to femtoliters. Overall, these advantages translate into intensely lower costs for DNA sequence production using NGS technologies (Shendure & Ji, 2008). Therefore, in whole-genome sequencing projects, the high-throughput NGS method may be favoured.

The major NGS technology platforms in the market for whole-genome sequencing are primarily from brand names such as Illumina, Roche 454, Solid and IonTorrent. Each platform has its advantages and disadvantages and cost implications in terms of reliability, time and money (Table 2.3). Depending on each NGS platform, they offer

values that may be attractive for specific purposes. For instance, the Ion Torrent is often positioned as a general purpose sequencer as well as in diagnostic protocols due to the quicker turnaround time (Tarabeux et al., 2014). However, longer reads technology offered by Illumina and Roche are desirable but the cost involved for Roche 454 FLX is very steep rendering it impractical for large scale genome projects. Pacbio reads are generally not used in large genome projects for direct sequencing but can be useful in resolving repetitive regions and ambiguous regions due to capability of generating very long read lengths (Table 2.3).

While there are many successful large genome sequencing studies (Cho et al., 2013; Daetwyler et al., 2014; Ge et al., 2013; R. Li et al., 2010), many whole-genome sequencing projects utilize the Illumina technology platform for reasons mainly being a high-throughput system and the lower cost involved when covering large amount of basepairs in large genomes (> 1Gbp). Due to the need of sequencing genomes at high sequencing coverage, the most cost-effective platform would be the high-throughput Illumina technology. Therefore, in this study, I have selected Illumina HiSeq for *de novo* sequencing the *M. javanica* genome at high sequencing depth because this genome is large and need high sequencing coverage for generating the first reference genome.

2.7.1 NGS process workflow

The common processes involved in general NGS platforms are library preparation, library amplification, and sequencing (Figure 2.3). The starting material for library preparation can be from either RNA or DNA (genomic source or PCR-amplified). In the case of RNA, it has to be transcribed into cDNA because as of now, NGS machines sequence only DNA directly. Since target library molecules sequenced on each NGS platform are required to be in specific lengths, genomic DNA requires fractionation and size selection, which is

performed by sonication, nebulization, or enzymatic techniques followed by gel electrophoresis and excision. For instance, Illumina NGs platform's standard fragment size is in the range of 300 and 550 bp including adapters. Generally, libraries are built by adding NGS platform-specific DNA adapters to the DNA molecules. These adaptors facilitate the binding of the library fragments to a surface such as a microbead (454, Ion PGM, SOLiD) or a glass slide (Illumina, SOLiD). With known adapter sequences, it allows the amplification of library fragments either by emulsion PCR (emPCR) or bridge PCR which are methods specific to the NGS platforms.

Table 2.3 Comparison between different NGS platforms

	454 FLX	**HiSeq2000**	**Solid 5500XL**	**IonTorrent (318 Chip)**	**PacBio**
Company	Roche	Illumina	Life Technologies	Life Technologies	Pacific Biosciences
Nucleotides/run	700 Mbp	540-600 Gbp	180 Gbp	800 Mbp	0.5 – 1 Gbp
Maximum Read length	700bp	2 x 100 bp	75+35 bp	200 bp	10kbp - >40kbp
Pairs	2x150 bp	2x150 bp	2x60 bp	N/A	
Run time	23 hours	11 days	12-16 days	4.5 hours	30m – 4 hours
Reagent cost per Mbp (USD)	7	0.04	0.07	1	0.13 - 0.60
Advantages	-Read length -Fast	-High throughput -Read length	-Low cost/base -Accuracy	-Less expensive -Fast	-Read length -Fast
Disadvantages	-Expensive -Homopolymer errors -Low throughput	-High DNA concentration -Short reads	-Slow -Palindromic errors -Short reads	-Homopolymer errors	-Error rate

In the sequencing step, NGS exploits parallelism in a factor of ten thousand to billions of library fragments. Generally, this is achieved by reiterated cycles of nucleotide addition using DNA polymerases or ligases (SOLiD), detection of incorporated nucleotides and washing steps. The extensive washing and repetitive steps, albeit automatic, may thus take sequencing to complete from hours to days. Basecalling is the process of algorithmically deciding the incorporated nucleotide from the signal intensities that are detected during sequencing process.

Every NGS platform relies on slightly different strategies to generate and detect signals (C. Luo, Tsementzi, Kyrpides, Read, & Konstantinidis, 2012; Merriman, R&D Team, & Rothberg, 2012; Ronaghi, Uhlén, & Nyrén, 1998). For instance, in case of Illumina and SOLiD sequencing, the four differently fluorescent-labelled nucleotides are flushed over the glass slide in parallel and emittance of fluorescence signals the combination of nucleotides being incorporated. In contrast, a sequential flooding of non-labelled native nucleotides occurs during 454 and Ion PGM sequencing where products of the enzymatic nucleotide incorporation reaction are detected by proton or pyrophosphate release. While proton release can be directly measured as pH change by the semiconductor chip of the Ion Torrent instruments (Merriman et al., 2012), the pyrophosphate signal is further converted into a light signal via subsequent reactions including the enzyme luciferase (Ronaghi et al., 1998). These signals, hence, will be converted into basecalls and converted into raw sequencing data that is generally output in FASTQ format that includes the basecall qualities along with the DNA sequence.

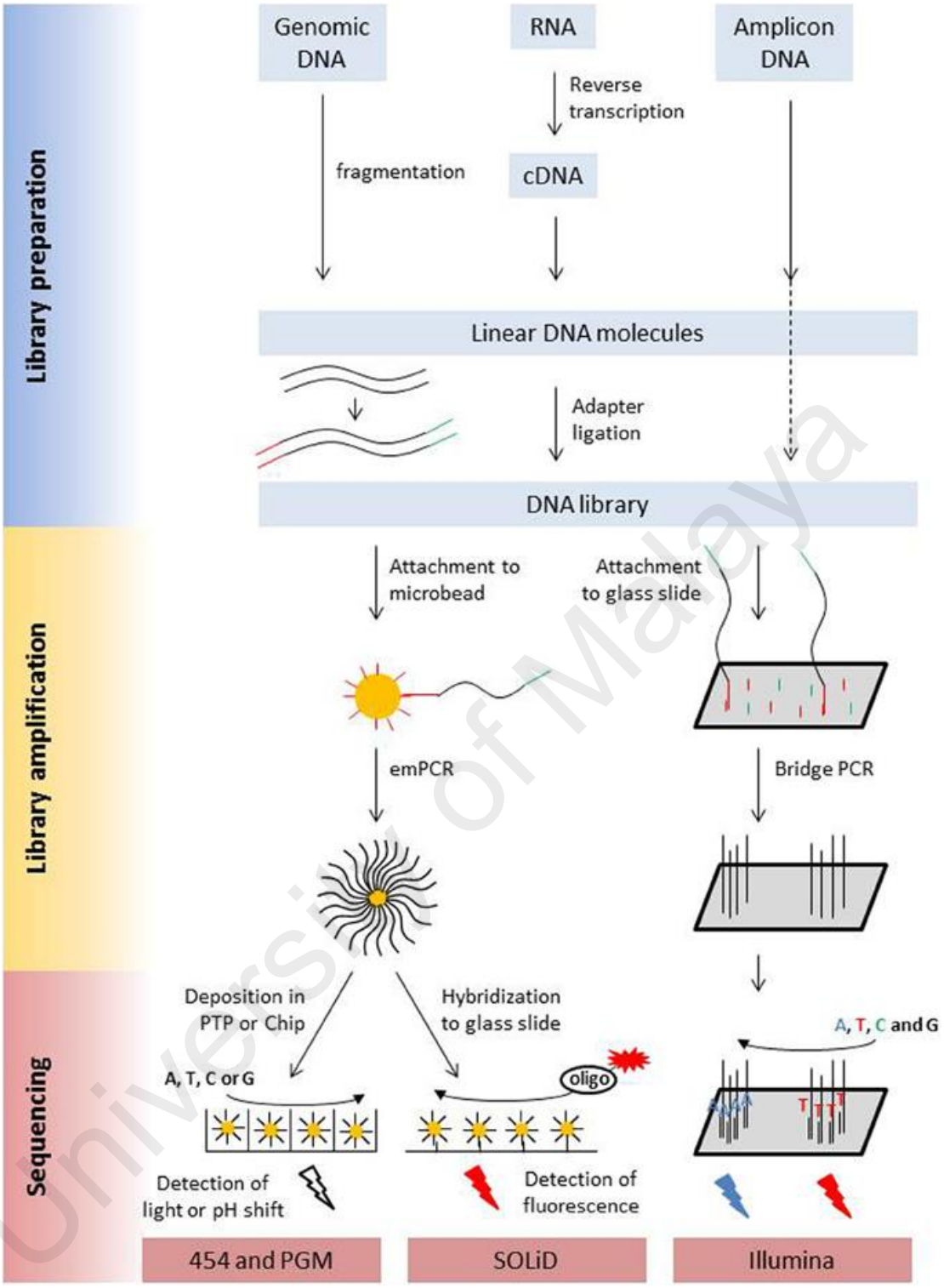

Figure 2.3 Schematic diagram of the process involved in common NGS platforms. **(Knief, 2014)**

2.7.2 NGS raw data pre-processing

As described earlier, NGS platforms suffer from a higher error rate compared to Sanger sequencing (Nakamura et al., 2011; Shendure & Ji, 2008). However, different approaches and algorithms can be developed to compensate and spot these errors (Margulies et al., 2005). As a simple approach, error-rates can be decreased by performing the DNA sequencing with high coverage, of at least 20 to 60-fold, depending on the sequencing project's goal (C. Luo et al., 2012; Margulies et al., 2005; Voelkerding, Dames, & Durtschi, 2009). Each sequencing read can be categorised as a distinct genotype or could be the result of sequencing error. Thus, it is very important to use established methods in differentiating these two of cause of variation as it may lead to inaccurate results when flawed.

To improve the quality of the data after base-calling, Phred-based filtering algorithms can be used to filter or remove low quality sequencing reads (Margulies et al., 2005). These filters discard reads with low-quality, uncalled, and ambiguous bases besides clipping the lower quality 3'-ends of reads. All such filters use the quality information contained in the FASTQ file that are computed by the NGS platform at each base during the base calling procedure. Previous studies (Minoche, Dohm, & Himmelbauer, 2011) have shown the effect of different filtering approaches on Illumina data and suggests that it can reduce error rates to less than 0.2% by eliminating around 15–20% of the low-quality bases, mostly via 3'-end trimming that are prone to errors. Another study has supported the findings that a 5-fold decrease of error rate can be observed by applying a filter (Phred score of Q30, with 0.1% likelihood of a false basecall) that eliminated reads with low quality bases (P. Nguyen et al., 2011). It may be useful to note that low quality bases are sometimes localised in specific regions of a genome. Removal of these reads may

introduce potential bias in the quantitative studies undertaken (Minoche et al., 2011; Nakamura et al., 2011).

Apart from read clipping and filtering methods, several error correction tools (e.g., Coral, HiTEC, Musket, Quake, RACER, Reptile, or SHREC) could be used as a complementary strategy to reduce sequencing error rates in reads (Knief, 2014). Generally, these error correction methods make use of high sequencing coverage in order to identify and correct errors using the laws of probability and statistics. Moreover, these algorithms often consider quality scores of the examined bases besides looking at neighbouring base quality values. For instance, some of these tools are able to correct substitution errors in Illumina sequencing data (Ilie and Molnar, 2013; Liu et al., 2013; Yang et al., 2013) while others (Coral, HSHREC, KEC, and ET) are designed to include indel correction algorithms that are available for the analysis of Roche's 454 and IonTorrent data (Salmela, 2010; Salmela and Schröder, 2011; Skums et al., 2012). The relevance of error correction tools is seen as a very useful strategy in *de novo* genome sequencing, resequencing and amplicon sequencing projects with benefits ranging from finding more optimal assembly in the DBG and reducing overall memory footprint to perform the assembly stage (Skums et al., 2012; Yang et al., 2013).

2.7.3 Genome assembly

After the pre-processing of sequence reads, we can assemble the pre-processed reads into contgis. The process of assembly of sequencing reads generated from NGS technology involves the reduction of redundant data by contiguously placing reads by overlapping them adjacent to each other in an optimal way (Miller, Koren, & Sutton, 2010). Instead of reads, when contigs (assembled set of reads) undergo the previously described process

with long length information, it is known as scaffolding. In other words, it is a process of reconstructing the target as such to groups reads into contigs and contigs into scaffolds.

Generally, the size and accuracy of the contigs and scaffolds are important statistics in genome assemblies (Miller et al., 2010). The quality of genome assemblies are usually described by maximum length, average length, combined total length, and N50. The contig N50 is the length of the smallest contig in the set that contains the fewest (of the largest) contigs whose joint length represents at least 50% of the assembly (Miller et al., 2010). Generally, larger N50 values imply a higher quality genome assembly that describes lesser overall fragments. Typical high-coverage genome projects have N50 values that range in megabases, however, they are dependent on the genome size and is not a good measure to compare between unrelated assemblies instead of the same. Assembly accuracy is tough to quantity. Nevertheless, mapping the assembled contigs/scaffolds to reference genomes is useful to examine its quality if the references exist.

As outlined earlier, an assembly is an ordered data construction that maps the sequencing data to a supposed reconstruction of the target (He, Zhang, Peng, Wu, & Wang, 2013). Contigs contain numerous sequence alignment of reads plus the consensus sequence. The scaffolds, at times called supercontigs or metacontigs, define the contig order and orientation and the sizes of the gaps between contigs. Scaffold topology may be a simple path or a network. Scaffold consensus sequence could have N's in the gaps between contigs. The number of consecutive N's may show the gap length estimate based on bridging paired ends (Miller et al., 2010).

There are many well-established software for assembling sequencing reads into contigs/scaffolds. In general, these genome assemblers can be grouped into three categories based on their approaches (Miller et al., 2010): (1) the Overlap/Layout/Consensus (OLC) approaches depend on an overlap graph; (2) the *de* Bruijn Graph (DBG) use some form of *k*-mer graph; and, (3) the greedy graph algorithms can use OLC or DBG (Table 2.4).

Table 2.4 Summary of assembly software used for *de novo* assembly of genomes

Assembler	**Algorithm**	**Preferred data**	**Multithreading**	**Target genome**
SGA	OLC	Illumina	Yes	Large
SOAPdenovo2	DBG	Illumina	Yes	Large
ALLPATHS-LG	DBG	Illumina	No	Large
CLC	DBG	Mixed	No	Large
SSAKE, SSHARCGS and VCAKE	Greedy	Illumina	No	Small
Edena	OLC	Illumina	No	Small
Newbler	OLC	454	No	Large
CABOG	OLC	Mixed	No	Large
Euler	DBG	454 + Sanger	No	Small
Velvet	DBG	Illumina	No	Small
ABySS	DBG	Illumina	Yes	Large

There are many factors that need to be considered when choosing the most appropriate genome assembler especially when considering for large whole genome sequencing project. Among these factors are the choice of algorithm, compatibility with the NGS platform, the support of the assembly of large genomes, and parallel-computing support for speeding-up the assembly (Zerbino & Birney, 2008). Generally, the choice of algorithm and software will directly determine the memory requirements and speed of assembly. Generally, DBG assemblers are faster but require large amount of memory compared to OLC assemblers.

2.7.3.1 Assessing the quality of the genome assemblies

After the genome assembly process, it is always useful to evaluate the quality of the genome assembly. It generally allows one to compare between different assembly efforts from different assemblers and parameters to identify the best possible assembly to use for further analysis. Previous studies from large collaborative efforts dubbed as Assemblathon 1 and Assemblathon 2 aimed at independently evaluating outputs from different assemblers (Bradnam et al., 2013; Earl et al., 2011). A study called Genome Asssembly Gold-standard Evaluations (GAGE) was similarly designed to critically evaluate several large-scale NGS projects (Salzberg et al., 2012). Lending from the wisdom of these evaluations, a straight-forward method called FRCbam was developed to quickly evaluate genome assembly without the often requisite reference genome (Vezzi, Narzisi, & Mishra, 2012). The FRCbam program calculates various types of suspicious features (Narzisi & Mishra, 2011) using the alignment file of reads against each contig in the assembly and outputs values that can be used to construct a Feature Response Curve (FRCurve) (Vezzi et al., 2012).

Apart from the genome quality metric, it is also important to estimate the completeness of the genome assembly. Due to possible coverage bias during sequencing certain regions may be under-represented by reads and subsequently may not be represented in the final genome assembly. Core Eukaryotic Genes Mapping Approach (CEGMA) is a tool to identify the universal core eukaryotic genes (CEG) that is found in the assembled genomes (Parra, Bradnam, & Korf, 2007). Depending on the percentage of CEG that were found, one can infer the gene completeness of the draft genome. High CEG completeness percentage relates to a highly complete genome assembly. Overall, the CEGMA and FRCbam tool are useful in evaluating genome assemblies apart from the basic genome assembly statistics.

2.7.4 Iterative assembly consideration for assembling mitogenomes

Assembling the mitochondrial genome from NGS reads of genomic DNA is a challenging process (Hahn, Bachmann, & Chevreux, 2013). It is difficult to distinguish nuclear or mitochondrial sequences in the NGS data since they are present as a heterogenous mixture during sequencing. The presence of nuclear copies of mitochondrial (Numt) DNA further hampers the effort by introducing misassemblies using common assembly software. Moreover, the high coverage if mitochondria worsens the signal-to-noise ratio which interferes with DBG based software. The DBG algorithm will not be able to function optimally when high coverage of target sequence is present as it will be overwhelmed by the abundance of the sequence.

To overcome these issues, an alternate assembly method such as an iterative assembly method using a closely related reference sequence is a better approach (Hahn et al., 2013; Hunter et al., 2015). Mitochondrial genomes of related organisms can be targeted as reference and used as a backbone, guiding the discovery and assembling of relevant reads from the pool of genomic reads where desired target is present (Hahn et al., 2013). In general, the reference DNA (fragments or genomes) is used to map the reads iteratively to extend the consensus sequence. In relation to mapping DNA sequences, the method enforces an inflexibility setting to limit the number of acceptable disparities in the sequence comparison to the reference preventing insertion of false positive reads that may cause issues in the downstream assembly process (Hahn et al., 2013). The iterative assembly strategy has been considerably successfully used to assemble organellar DNA successfully in the past using NGS data (Alam, Petit III, Read, & Dove, 2014; Austin, Tan, Croft, & Gan, 2016; Bagatharia et al., 2013; Chen et al., 2014).

2.8 Genome annotation

After the generation of genome assemblies, the next natural step is to understand the unknown sequences through the process of genome annotation in order to identify the functional elements in the genome assemblies. Genome annotation is the procedure of conferring biological information to sequences. It usually consists of two main stages: recognizing elements on the genome, usually called gene prediction, and attaching biological/functional information to these elements.

Repeat masking is the first step in annotating the repetitive regions in the genome often replete with a variety of transposable elements. By nomenclature, these transposons are subdivided into two main classes called the Class I retrotransposons and the Class II DNA transposons. Class I elements are further divided into four main classes that are LTR, DIRS, PLE, LINE and SINE. Class II elements are divided into TIR, Crypton, Helitron and Maverick (Wicker et al., 2007). To generate proper gene predictions for the target genome, masking the repeats using well-established RepearMasker software (Smit, Hubley, & Green, 1996) is an important step as to remove interference in applying the available gene models by gene prediction software.

Structural annotation encompasses of the identification of genomic elements: transposable elements, open-reading frames (ORFs) and their localized gene structure, coding/non-coding regions, and the location of regulatory motifs (Yandell & Ence, 2012). On the other hand, functional annotation entails attaching biological information to genomic elements: biochemical function, biological function, involved regulation and interactions, and expression.

There are many bioinformatics software that have been developed to identify and predict genes in the genomes (Table 2.5). Among these popular software are Genemark,

HMMER, BLAST and Genscan (Borodovsky & McIninch, 1993; Burge & Karlin, 1997; Finn, Clements, &Eddy, 2011; Johnson et al., 2008). Nevertheless, to systematically perform the genome annotation, some pipelines have been developed to combine best annotation practices into a single software which are carried out through a pipeline associating multiple types of software for automatic execution. For instance, MAKER is one of the most commonly used genome annotation pipeline in many large genome sequencing projects. This pipeline executes repeat masking, gene and tRNA identification in a single workflow (Cantarel et al., 2008).

Table 2.5 List of bioinformatics tools for gene and structural annotation of genomes

Software	Description
***Ab initio* and evidence-based gene finders**	
Augustus	Accepts expressed sequence tag (EST)-based and protein-based evidence hints. Highly accurate
mGene	Support vector machine (SVM)-based discriminative gene predictor. Directly predicts 5′ and 3′ untranslated regions (UTRs) and poly(A) sites
SNAP	Accepts EST and protein-based evidence hints. Easily trained
FGENESH	Training files are constructed by SoftBerry and supplied to users
Geneid	First published in 1992 and revised in 2000. Accepts external hints from EST and protein-based evidence
Genemark	A self-training gene finder
Twinscan	Extension of the popular Genscan algorithm that can use homology between two genomes to guide gene prediction
GAZE	Highly configurable gene predictor
GenomeScan	Extension of the popular Genscan algorithm that can use BLASTX searches to guide gene prediction
Conrad	Discriminative gene predictor that uses conditional random fields (CRFs)
Contrast	Discriminative gene predictor that uses both SVMs and CRFs
CRAIG	Discriminative gene predictor that uses CRFs
Gnomon	Hidden Markov model (HMM) tool based on Genscan that uses EST and protein alignments to guide gene prediction
GeneSeqer	A tool for identifying potential exon–intron structure in precursor mRNAs (pre-mRNAs) by splice site prediction and spliced alignment

Software	Description
Aligners	
BLAST	Suite of rapid database search tools that uses Karlin–Altschul statistics
BLAT	Faster than BLAST but has fewer features
Splign	Splice-aware tool designed to align cDNA to genomic sequence
Prosplign	Global alignment tool that uses BLAST hits to align in a splice-site- and paralogy-aware manner
Exonerate	Splice-site-aware alignment algorithm that can align both protein and EST sequences to a genome
MapSplice	Spliced aligner that does not use a model of canonical splice junction
TopHat	Transcriptome aligner that aligns RNA sequencing (RNA-seq) reads to a reference genome using Bowtie to identify splice sites
GSNAP	A fast short-read assembler
JIGSAW	Combines evidence from alignment and *ab initio* gene prediction tools to produce a consensus gene model
EVidenceModeler	Produces a consensus gene model by combining evidence from protein and transcript alignments together with *ab initio* predictions using weights for both abundance and the sources of the evidence
GLEAN	Tool for creating consensus gene lists by integrating gene evidence through latent class analysis
Evigan	Probabilistic evidence combiner that use a Bayeisan network to weigh and integrate evidence from *ab initio* predictors, alignments and expression data to produce a consensus gene model
Genome annotation pipeline	
PASA	Annotation pipeline that aligns EST and protein sequences to the genome and produces evidence-driven consensus gene models
MAKER	Annotation pipeline that uses BLAST and exonerate to align protein and EST sequences. Also accepts features from RNA-seq alignment tools (such as TopHat). Massively parallel
NCBI	The genome annotation pipeline from the US National Center for Biotechnology Information. Uses BLAST alignments together with predictions from Gnomon and GenomeScan to produce gene models
Ensembl	Ensembl's genome annotation pipeline. Uses species-specific and cross-species alignments to build gene models. Also annotates non-coding RNAs

* Yandell & Ence (2012)

2.8.1 MAKER annotation pipeline

Traditionally, bionformaticians try to understand an unknown DNA sequence by sequence similarity searches against established and manually annotated databases. Later, researchers developed methods of predicting genes from DNA. Although expert manual curation and annotation is ideal, however, it can be time consuming and repetitive when involving large genomes. The recent developments in large genome annotation is to automate the process as much as possible while providing important information for expert annotators to edit thereafter. One example of the most commonly used annotation automation is by the automated genome annotation pipeline called MAKER (Cantarel et al., 2008). The repetitive tasks of repeat masking, aligning protein and RNA evidence, *ab initio* gene finding, and scoring the ensemble evidence is packaged into the MAKER pipeline in a highly parallelized system suitable for grid or cluster computing.

The outline of the MAKER pipeline is given in Figure 2.4. In general, the workflow begins by masking the repeat regions using RepeatMasker followed by alignment of known proteins and available EST/mRNA sequences to the genome. The resulting alignments are then compared with the Exonerate, a pairwise sequence comparison software. Additionally, when configured, MAKER will use SNAP, Genemark and Augustus *ab initio* gene prediction software to generate gene predictions. In the final step, MAKER can utilize the prior evidence from alignments and multiple gene predictions to score and produce the best gene model for the genome.

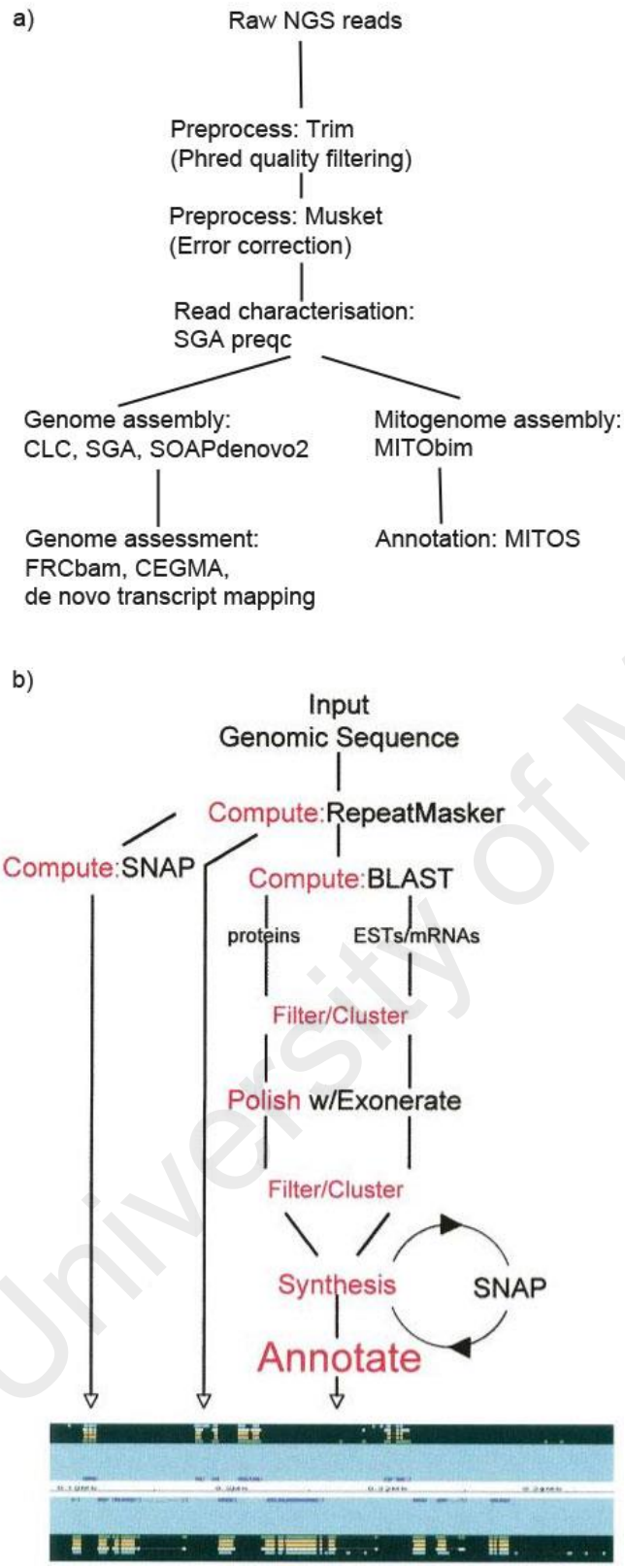

Figure 2.4 a) Typical NGS data pre-processing and assembly process b) MAKER annotation workflow. All the tools outlined in a) and b) are used in this study.

The data raw NGS data involves the pre-processing, assembly and annotation steps. Annotation process can be automated using a pipeline such as MAKER. The MAKER annotation pipeline aggregates various evidence from BLAST searches, EST/mRNAs and *ab initio* as well as trained gene prediction software to generate the gene model.

2.9 The lncRNA repertoire

Non-coding-RNA (ncRNA) genes are genes that do not code for any known proteins (Quinn & Chang, 2016). These genes were once thought to be limited to ribosomal, transfer, spliceosomal, and other essential RNAs, but some emerging studies have shown it to be far more diverse (Fatica & Bozzoni, 2014). These non-canonical class of ncRNA genes are categorised into short ncRNAs (i.e. microRNAs, snoRNAs) and long ncRNAs (lncRNAs). While importance of microRNAs as post-transcriptional repressors is known, however, there is a lack of understanding particularly on lncRNA genes (Carninci and Hayashizaki 2007). Besides being order of magnitude more pervasive than microRNAs, lncRNA are typically longer than 200 bp and may be expressed either in single or multi-exonic form.

The identification and study of lncRNAs is of major relevance to human biology and disease since they represent an extensive, largely unexplored, and functional component of the genome (Mattick 2009; Ponting et al. 2009). Since pangolins exhibit a wide range of adaptations to their unique characteristic, it may be useful to compare the identified lncRNA in pangolins to expand and validate known lncRNA annotations in humans and other organisms. For instance, although some lncRNAs have been shown to affect in human diseases, these studies are limited by the lack of lncRNA annotations. Therefore, the identification of high-quality catalogues of lncRNAs and its expression in tissues is important work. On the contrary, similar information for protein-coding genes has long been obtainable.

In order to facilitate the growing interest in ncRNA and particularly lncRNA, generating high-confidence catalogues of lncRNA for further study is crucial. Genome sequencing

efforts in association with RNA-seq in several studies were able to catalogue lncRNA of various genomes of the sponge, mouse, bovine, and human. These studies have shown that a large portion of these genomes may be actually transcribed and could aid the discovery of many more non-coding transcripts (Carninci et al. 2005; Harrow et al. 2006). Furthermore, some studies have focused on large intergenic non-coding RNAs (lincRNAs) (Ponjavic et al. 2007; Guttman et al. 2009; Khalil et al. 2009; Orom et al. 2010), the non-coding transcripts which are away from protein-coding regions.

Recent advances in whole-transcriptome RNA sequencing (RNA-seq) (Mortazavi et al. 2008) and computational methods for transcriptome reconstruction now provide an opportunity to expansively annotate and characterize lncRNA transcripts (Guttman et al. 2010; Trapnell et al. 2010; Garber et al. 2011). Certainly, an initial application of this method in three mouse cell types characterized the gene structure of more than a thousand mouse lincRNAs, most of which were not identified before (Guttman et al. 2010). A few dozen lncRNAs are known to play important regulatory roles in diverse processes, such as X inactivation (XIST) (Zhao et al. 2008), imprinting (H19 and KCNQ1OT1) (Leighton et al. 1995; Pandey et al. 2008), and development (HOTAIR and COLDAIR) (Rinn et al. 2007; Heo and Sung 2011).

With logical frameworks that have been thoroughly established in previous studies, it is now possible to catalogue and curate high-confidence lncRNA. Generally, a well-designed filtering pipeline can be used to arrive at putative lncRNA that can be further assessed for protein-coding capacity. The final set that exhibit poor protein-coding potential will remain as high-confidence lncRNA.

CHAPTER 3: MATERIALS AND METHODS

The procedures used in the present study, especially tissue sampling and storage were based on previously published guidelines for undertaking genome projects (Wong et al., 2012).

3.1 Retrieval of genome and raw reads data

As part of the comparative genomics approach undertaken in this study, the whole-genome sequencing data (raw reads and draft assembly) of *M. pentadactyla* genome sequence were obtained from our collaborators Wesley Warrens from the University of Washington. The sequence was available at NCBI under the Bioproject accession number PRJNA20331, and the assembled genome was available through accession number JPTV00000000.1.

3.2 Animal sampling

For genome sequencing, a female pangolin (*M. javanica*) were used in this study. The animal was provided by the Department of Wildlife and National Parks Malaysia (DWNPM) under Special Permit No. 003079 (KPM 49) for endangered animals. All animal studies performed were approved by the International Animal Care and Use Committee (IACUC) of University of Malaya (DRTU/11/10/2013/RH (R)).

3.2.1 Animal handling

Animal handling was performed with assistance of veterinary officers from the Department of Wildlife and National Parks, Malaysia (DWNPM). Prior to organ harvesting and blood sampling, the pangolins were injected with the anaesthetic ketamine-xylazine 1:1 mixture (dosage: 0.5-1mL for each pangolin) or alternatively with

Zoletil (3-4mg/kg) via intramuscular (IM) for minimal intervention. When muscles have shown signs of relaxation, blood sampling was performed using a 5mL syringe and 21G needle via the coccygeal vein found at the tail.

For the pangolin genome sequencing, the female is preferred over the male organism. This is to obtain equal coverage of the X chromosome as compared to the autosomes, to ensure similar quality for the whole-genome shotgun assembly of both autosomes and chromosome X. Chromosome Y is not amenable to sequencing strategy due to its large palindromic regions, whereas the X chromosome is much more stable and retains most of its autosomal characteristics. One female pangolin weighing 4.73kg labelled UM3 was used and specifically liver, cerebrum and cerebellum tissue were utilized for genome sequencing.

3.2.2 Harvesting of tissue samples from organs

After blood sampling, animals were euthanized by DWNPM veterinarians. The procedure involves administration of Dolethal® (Pentobarbitone sodium 200mg/L; dosage: 1mL/kg body weight). Several organs were harvested and used in this study including the cerebrum, cerebellum and liver. Organs were harvested only when the animal was pronounced dead by the veterinary surgeon upon checking for vital signs, i.e. breathing and pulse which stopped within 5 minutes of administration of the drug. Organ harvest and other sampling was performed within 15 minutes of dissection. No experiments were done on live animals.

3.3 DNA extraction

M. javanica DNA was extracted from the liver, cerebrum and cerebellum tissue samples of the female UM3 pangolin while DNA from blood were extracted from three female

(UM3, UM2 and UM2) and three male pangolins (2T9, 12T, and 2T2) using a Qiagen 20/G genomic tip following manufacturer's protocol, as outlined below. For this study, high-molecular weight (HMW) DNA was extracted from blood, liver, cerebrum and cerebellum.

For DNA extraction from blood, both whole blood and white blood cells (buffy coat) isolated by centrifugation were used. For either case protocol was started with 1 ml of whole blood or buffy-coat suspension and added ice-cold Buffer C1 and 3 ml of ice-cold distilled water. The tube was inverted several times until the solution became translucent and was then immediately placed on ice for 10 minutes for the cell lysis step. The tube of lysed blood was centrifuged at 1300 x g for 15 minutes at 4°C and its supernatant discarded thereafter. Ice-cold Buffer C1 (0.25 ml) and distilled water (0.75 ml) each were added and subjected to resuspension by vortexing. The mixture was then centrifuged again at 1300 x g for 15 minutes at 4°C and its supernatant discarded. Subsequently, 1 ml of Buffer G2 was added along with 0.1 ml of Proteinase K followed by vortex mixing and incubation at 50°C for 60 minutes in a water bath.

For cerebrum, cerebellum and liver tissues, a Qiagen TissueRupter was used to homogenize approximately 10 mg of the tissue with 0.5 ml of Buffer G2. Subsequently, 1.5 ml of Buffer G2 was added along with 0.1 ml of Proteinase K followed by vortex mixing and incubation at 50°C for 60 minutes in a water bath.

The next steps are the same for both products from blood and liver lysis and enzymatic digestion steps. The Qiagen Genomic-tip 20/G were equilibrated by passing through Buffer QBT through its column and emptying it by gravity flow. The sample from previous step, the lysed and digested blood or tissue homogenate (of cerebrum,

cerebellum and liver), were then vortex mixed and applied through the column and emptied by gravity flow. When samples were flowing too slowly, some positive pressure was applied to the column. Once the columns were emptied, the columns were washed thrice using 1 ml of Buffer QC. Next, elution was performed by adding slightly warmed eluent under 50°C bath, Buffer QF to the column and collecting the eluate. Finally, in order to precipitate the DNA, 1.4 ml of isopropanol was added and mixed. The mixture was centrifuged immediately at 5000 x g for at least 15 minutes at 4°C. The DNA pellet was washed with 1 ml of 70% ethanol vortex mixed and centrifuged at >5000 x g for 10 minutes at 4°C. The supernatant was carefully pipetted out and the tube was air-dried for 20 minutes. The DNA was then reconstituted with 100 µl Tris-EDTA buffer (pH 8) and was allowed to sit overnight prior to -80°C storage.

3.3.1 Sample quality assessment and quantification

The extracted DNA were checked for quality in terms of UV absorbance ratio at various wavelengths of 260nm/280nm and 260nm/230nm using Nanodrop 2000 UV-Vis spectrophotometer where high-purity DNA exhibits ratios close to 1.8 and 2.0 respectively. The values indicate high quality and represent negligible contamination for further downstream processing. The concentration of the DNA was estimated using blank Tris-EDTA buffer as a control. In order to further assess the integrity, a visual inspection of DNA bands using gel electrophoresis was performed on a 1% agarose gel at 60V for 45 minutes.

3.3.2 Library preparation and whole-genome shotgun sequencing

DNA Libraries for the *M. javanica* were constructed as per sequencing protocol for Illumina HiSeq 2000. The insert sizes of the libraries that were constructed were 180 bp, 500 bp, and 800 bp which were short insert sized paired-end libraries, while insert sizes

of 2 kb, and 5 kb were long insert sized mate-pair libraries. The insert size was constructed with the genome assembly process in mind. Based on ALLPATHS-LG, one of the more sophisticated genome assemblers, 180bp library of 100bp length is one of the prerequisites due to its capability of the reads pair to overlap by 20bp. The other insert libraries were also chosen based on the assembler's recommendation. An increasing size of insert library such as 500bp and 800bp will assist the assembly process while longer mate pair libraries at 2kbp and 5kbp can be used for the scaffolding step. The libraries were constructed and sequenced using Illumina HiSeq2000 with 100bp paired-end strategy at an estimated sequencing coverage of approximately 100-fold at BGI, Hong Kong.

3.3.3 Data pre-processing and error correction

The quality of the sequencing reads was evaluated by FASTQC tool (Andrews, Krueger, Seconds-Pichon, Biggins, & Wingett, 2014). The sequencing reads were filtered using PRINSEQ with a quality threshold of Phred 20 score. The pre-processed reads were then error-corrected using MUSKET 1.1 (Y. Liu, Schröder, & Schmidt, 2013). The error-corrected reads were subsequently used for various downstream analysis including genome assembly.

3.3.4 Genome size estimation

3.3.4.1 Genome size estimation

To estimate the genome size, *k*-mer was counted using the tool Preqc available from the String Graph Assembler (SGA) genome assembler (Simpson & Durbin, 2010). The tool counted and plotted 31-mer histogram from 20,000 reads and estimated the genome size by identifying the peak of the Poisson distribution. The estimation is based on the

principle as follows, where the mean number of times a unique genomic *k*-mer appears in the reads, λ_k is as follows:

$$\lambda_k = \frac{n(l-k+1)}{G}$$

Where n is the number of reads, l is the read length and G being the genome size.

3.4 Genome assembly

3.4.1 Whole-genome assembly

The genome assemblies were performed using three different assemblers: CLC Assembly Cell 4.10 (CLC bio, Aarhus, Denmark), SGA 0.10.10 (Simpson & Durbin, 2010) and SOAPdenovo2 (R. Luo et al., 2012) and the best assembly was chosen for downstream analyses in this whole study. These assemblers generally utilize graph theory principles to aid assembly. For instance, SOAPdenovo2 utilizes the DBG algorithm for contig assembly while SGA uses string graph for assembly which was earlier described by Eugene Myers (Myers, 2005), but further optimized by the Simpson-Durbin algorithm by utilizing the FM-index (Simpson & Durbin, 2010).

Genome assembly was also explored with other tools which subsequently aided in finalizing the three assemblers of choice for further evaluation. All the assemblies using CLC, SOAPdenovo2 and SGA were performed by exploring select parameters respective to the assemblers. An important parameter, the *k*-mer size was explored for CLC (33, 43, 51 and 63) and SOAPdenovo2 (odd numbers from 29 – 55). For SGA, the overlap length parameter was varied from 75 to 95 at odd numbers.

3.4.1.1 Comparing assemblies with Feature Response Curve

A candidate genome assembly was picked from the three assemblers each based on the highest N50 metric calculated using QUAST(Gurevich, Saveliev, Vyahhi, & Tesler,

2013). The Feature Response Curve (FRC) was produced using the FRCbam program (Vezzi et. al, 2012). The raw sequencing reads were mapped using Bowtie2 (Langmead & Salzberg, 2012) to each candidate genome and the cumulative feature density were plotted for each genome assembled using CLC, SOAPdenovo2 and SGA. The higher the number of features per cumulative nucleotide size, the assembly was regarded to be better. In this way, the best assembly was selected for subsequent scaffolding step.

3.4.1.2 Scaffolding and gap-closing

To extend the contigs into longer fragment/scaffolds, long range size-selected mate pair libraries were used to provide positional information for scaffolding the contig sequences. Thus, the scaffolding performed using SOAPdenovo2 requires mapping both short reads and long reads onto the assembled contigs and placing these contigs into longer scaffolds often separated by ambiguous N sequences called gaps. These gaps were filled by the program wherever possible using the mapped read information.

3.4.2 Quality assessment of genome assembly

While the Feature Response curve has provided with the quality of the assembly of outputs from different assembly software, the chosen best final assembly of the genome was further assessed using standard N50 metric, transcript mapping and CEGMA analysis. More information is provided in the following subsections.

3.4.2.1 RNA-seq transcripts mapping

Independently *de novo* assembled transcripts from Trinity (Grabherr et al., 2011) as well as Unigenes that have been stringently screened using multiple transcript assessments were used to validate the draft genome assemblies. A total of 89,754 consensus transcripts/UniGenes generated by combining transcriptomic fragments were used from

three different assemblers in a separate study of *M. javanica* transcriptome project (Aini et al., 2016). Briefly, the transcriptomes of eight pangolin organs including cerebellum, cerebrum, lung, heart, kidney, liver, spleen and thymus were sequenced using Illumina HiSeq 2000 technology platform (2x100bp strategy). The pooled sequencing reads from the eight samples were *de novo* assembled using the three assemblers independently: Trinity (Grabherr et al., 2011), SOAPdenovo2 (R. Luo et al., 2012) and Velvet (Zerbino & Birney, 2008). The assembled transcriptomic fragments/transcripts were clustered using CD-Hit-EST program (Huang, Niu, Gao, Fu, & Li, 2010) with a clustering threshold of 98% sequence identity. The longest sequence representatives in each clustered transcript were selected and classified as the UniGenes. A total of 89,754 consensus UniGenes (the transcript clusters that have transcripts generated from the three different assemblers) was used for this analysis. Apart from Unigenes, Trinity generated transcripts for alignment were also used. Splice-junction aware mapper such as GMAP (T. D. Wu & Watanabe, 2005) was used to map the transcripts to the draft pangolin assembly in order to identify the mapping rate.

3.4.2.2 CEGMA analysis

The genome assembly quality of *M. javanica* and *M. pentadactyla* were evaluated with the CEGMA pipeline. CEGMA is a computational method that relies on a defined set of ultra-conserved eukaryotic protein families for building a highly reliable set of gene annotations. By looking at core eukaryotic gene present as both complete and partial hits in the draft genome, the percentage of these identified core genes can be used to infer the quality of our final draft assembly.

3.5 Evolutionary analysis

3.5.1 Speciation and divergence time

To estimate divergence time of Pholidota from Carnivora lineage, MCMCTREE software as part of PAML 4.8. package (Z. Yang, 2007) was used. Single copy gene families were used to construct a phylogenetic tree for both *M. javanica* and *M. pentadactyla* and other closely related mammalian species. Four-fold degenerate sites were extracted from the alignment of 1423 1:1 orthologs of 17 species resulting in 107,351 sites. Totally seven reliable calibration points from the Fossil Calibration Database (Ksepka et al., 2015) were used as priors (Table 3.1).

Two runs of MCMCTREE with 3,000,000 generations and 300,000 burn-in were conducted and convergence of both runs was checked using Tracer 1.5 software (Rambaut, Suchard, Xie, & Drummond, 2014). Tree was visualised using FigTree software (Figure 4.14**)** (Rambaut, 2014).

Table 3.1 Calibration points for the divergence time estimation.

No.	Taxa	Clades	Minimum bound	Maximum bound	Method	Reference
1	Human - mouse	Archonta-Glires	61,5	100,5	biostratigraphy	(M. J. Benton & Donoghue, 2007)
2	Primates+mouse - dog, horse, cow	Euarchontaglires -Laurasiatheria	61,6	164,6	fossil	(M. Benton & others, 2015; Ksepka et al., 2015)
3	Cat-dog	Carnivora	37,3	66	fossil	(M. Benton & others, 2015; Ksepka et al., 2015)
4	Megabat-Microbat	Chiroptera	45	58,9	fossil	(Ksepka et al., 2015; Phillips, 2015)
5	Horse+rhino - pig+cow.	Common ancestor of Cetartiodactyla	52,4	66	fossil	(M. Benton & others, 2015; Ksepka et al., 2015)

6	Mouse-rat	Mus - Rattus	10,4	14	biostratigraphy	(M. Benton & others, 2015; Ksepka et al., 2015)
7	Hedgehog-Common shrew	Lypotyphla	61,6	164,6	fossil	(M. Benton & others, 2015; Ksepka et al., 2015)

3.5.2 Ancestral population size estimation

Utilizing the Pairwise Sequential Markovian Coalescent (PSMC) model tool previously developed in another study (H. Li & Durbin, 2011), the historical effective population size (N_e) over time were inferred for both the pangolin genomes. The short insert read libraries were mapped to assembly to produce a BAM alignment file using Bowtie2 (Langmead & Salzberg, 2012). Then, the PSMC script was used to generate the consensus diploid genome as per variant-calling protocol, that infers the population size history and to perform a bootstrap 100 times (H. Li & Durbin, 2011). In order to scale the population size history, N_e, a generation time, g of 7 years and per site per generation mutation rate of 1.4×10^{-8} was used which was calculated based on neutral theory as per computed phylogenetic tree and divergence time in the previous section.

3.6 Genome annotation

3.6.1 Transposable elements annotation

3.6.1.1 Known TE identification

Known transposable elements (TE) were identified using RepeatMasker open-4.0.5 (Smit, Hubley, & Green, 2013) against the Repbase TE database (version 2014-01-31). The program was run with species parameter set to mammal. The repeat landscape across Kimura distance was identified using RepeatMasker scripts.

3.6.1.2 *De novo* repeats

A pangolin specific repeat library was compiled *de novo* using RepeatModeler (Smit & Hubley, 2013). The results were a set of consensus sequences and putative classification information for the identified repeats. Subsequently, the consensus sequence was used as a library to identify repeats *de novo* using RepeatMasker.

3.6.1.3 Segmental duplications

Segmental duplication (SD) sets for *M. javanica* and *M. pentadactyla* were constructed using a modified WGAC method (Gokcumen et al., 2013). LAST (Kielbasa et al., 2011) was used to perform alignment of the draft genome against itself. Matches that lie in regions of high-copy repeats annotated by RepeatMasker were filtered. These filtered matches were then extended using the clasp program (Otto, Hoffmann, Gorodkin, & Stadler, 2011). Subsequently, the resulting chained alignments were filtered if less than 1000 bp. The remaining chains were globally aligned using either stretcher or needle from the European Molecular Biology Open Software Suite (EMBOSS) package, depending on the chain size (i.e., when the product of the sequence lengths was greater than 100 Mb stretcher was used, otherwise needle was used). Alignments of smaller than 90% identity, or a gap percentage larger than 30% were discarded.

3.6.2 Structural annotation

Structural annotation of transposable elements, gene and tRNA sequences were performed using the MAKER pipeline software (Cantarel et al., 2008). The program was run using default parameters.

3.6.2.1 Comparative repeat composition of genic, intronic and intergenic regions

Available Carnivoran genome was downloaded from NCBI database along with the GFF annotation file. Using bedtools, structural regions such as genic, intronic and intergenic coordinates were calculated with the GFF file (Quinlan & Hall, 2010). Repeatmasker was rerun on the downloaded genome. Repeats were labelled corresponding to the structural region overlaps of at least one basepair was present in the repeat to either genic, intronic, and intergenic.

3.7 Mitogenome assembly and annotation

3.7.1 Mitogenome assembly

For the mitogenome assembly of the *M. javanica*, a 180bp paired-end library containing 804,047,380 reads were used. In comparison, for *M. pentadactyla*, publicly available NGS reads from the NCBI Sequence Read Archive (SRA) accession SRR770301 containing 3,240,714 reads were used to assemble the mitogenome separately. MITObim v 1.6 (Hahn et al., 2013) were run alongside MIRA 3.4.1.1 (Chevreux, 2007) to iteratively map and extend the assembly of the mitogenome using a reference pangolin mitogenome (NCBI accession: NC_016008)(Qin et al., 2012). The resulting assembly for each of the mitogenome were trimmed based on overlaps of the ends as present in a circular DNA.

3.7.2 Mitogenome Annotation and tRNA Secondary Structure Prediction

All mitogenome assemblies were annotated with MITOS webserver revision 656 (Bernt et al., 2013). The annotated tRNA genes were extracted for each species and putative folding was predicted using RNAfold program from the Vienna package 2.1.9 (Hofacker, 2009) with structural constraints (-C), as derived from MITOS annotations.

3.7.3 Alignment of the mitogenomes

The assembly of the mitogenome were assessed using whole mitogenome alignments with other complete mitogenome sequences of *Manis* sp. found in the National Centre of Biotechnology Institute (NCBI) database.

3.7.4 Taxonomic placement of assembled genome using the mitogenome

To study the phylogenetic relationships of pangolin species, a phylogenetic tree using Cytochrome B (*CYTB*) gene sequences were reconstructed from this project. Three *CYTB* gene sequences representing wild *M. javanica* from the database of the DWNPM and also from the public NCBI database.

All *CYTB* nucleotide sequences were aligned using MUSCLE (Edgar, 2004). Prior to analysis, the best model which describes the base substitution was calculated. The resulting Hasegawa-Kishino-Yano model considering some sites to be evolutionarily invariable (HKYI) was used along with Maximum Composite Likelihood approach to estimate pairwise distances (Hasegawa, Kishino, & Yano, 1985). Upon ignoring positions containing gaps and missing data, there were a total of 296 positions in the final dataset of 49 sequences. The phylogenetic tree was constructed based on bootstrap values of 1000 replicates (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013).

To construct a more robust tree, a phylogenetic tree using the whole mitogenome sequences was constructed. The full-length mitogenome sequences were aligned using MAFFT v7.113b and imported into MEGA6 for tree reconstruction (Katoh, Kuma, Toh, & Miyata, 2005); (Tamura et al., 2013). Prior to analysis, the best model which describes the base substitution was calculated. The resulting General Time Reversible with a discrete Gamma distribution model (GTR+G) was used along with Maximum Composite

Likelihood approach to estimate pairwise distances. Upon ignoring positions containing gaps and missing data, there were a total of 16,457 positions in the final dataset of 8 mitogenome sequences. The phylogenetic tree was constructed based on bootstrap values of 1000 replicates (Tamura et al., 2013).

3.7.5 Codon usage patterns of the pangolin mitogenome

Codon usage, relative synonymous codon usage (RCSU), and GC3 were calculated using INCA for the entire set of protein-coding genes (PCG) (*COB*, *COX1*, *COX2*, *COX3*, *NAD1*, *NAD2*, *NAD3*, *NAD4*, *NAD4L*, *NAD5*, *NAD6*, *ATP8* and *ATP6*) and the concatenated set of protein coding genes (PCG) (Supek & Vlahoviček, 2004). The codon usage and RSCU values were clustered by species respectively and heat maps were plotted using R statistical package for each gene (Fischer et al., 2013).

3.8 Presence of bacteria in *M. javanica*

3.8.1 Discovery of bacterial sequences in pangolin whole-genome data

Following the indication of foreign DNA presence during read characterization (Figure 4.4 c), the *M. javanica* genome was subjected to contamination screening. Next, tissue-specific genome assemblies were generated with CLC Assembly Cell using NGS sequencing data from cerebrum and cerebellum samples, which were used to screen for bacteria using BLASTN package. Bacterial nucleotide sequence database was used as reference for BLASTN, together with stringent threshold (90% identity and 90% coverage) to reduce false positive hits.

3.8.2 Tissues subjected to testing

Apart from genomic evidence of the presence of bacterial reads in pangolin genomic reads, tissue samples of *M. javanica* were also tested for the same. We screened cerebrum, cerebellum, liver, lungs, heart, spleen, thymus and skin tissues using polymerase chain reaction (PCR) assays with our in-house designed primers. All the organ material was harvested within 15 minutes and was snap frozen for subsequent storage in -80 C. For DNA extraction, minimal thawing was performed to obtain the tissue samples.

3.8.3 *B. fungorum* screening using PCR assays

PCR assays were performed with assistance of a colleague, Tan Ze King. Three different target genomic regions as identified from previous BLASTN results, that showed top hits to the bacteria were selected for primer design and PCR. Primer sequences used are shown in Table 3.2. *Burkholderia*-specific primers pairs targeting 2739bp Burkholderial transposase genomic region (90% identity and 90% coverage nucleotide homology) was Target A. *Burkholderia*-specific primers pairs targeting 957bp of OI25_7129 hypothetical protein (95% identity and 100% coverage nucleotide homology) was Target B. Whereas specific primer pair targeting 4655 bp of *Burkholderia* DNA polymerase genomic region (100% identity and 100% coverage) was Target C. The bacterial universal 16S primer pairs targeting approximate 1500 bps was used as control.

The primers were used to amplify the fragment for subsequent PCR as per the following protocol. The total reaction volume of 50 µL contained 160ng purified organ gDNA, 0.3 mol of each primer, deoxynucleotides triphosphates (dNTP, 400 µM each), 1.0 U *Taq* DNA polymerase and supplied buffer were used. The PCR was performed using the following protocol: 1 cycle (94 ºC for 2 minutes) for initial denaturation; 35 cycles (98 ºC for 10 secs; 68 ºC for 3 min) for annealing and DNA amplification. The PCR products

were purified by the standard methods and directly sequenced with the same primers using BigDye© Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems).

Table 3.2 PCR primers designed for screening *Burkholderia fungorum*

Gene	Primers	Length (bp)	Annealing Temperature
Target A	5'- CATCTTCGCCTTCTTGACGTTCTC -3' 3'- GAATAAATGAAGCGTCCCAGAGACG -5'	2739	68 °C
Target B	5'- GATACGTTGTCTGCGCTGGGCACC -3' 3'- CGTCTTTCTCACGGTGTTTCAGAG -5'	957	72 °C
Target C	5'- GAATGTGACCATGGCGGCACCGGTCGCCG ACCAC -3' 3'- CCGGGTGCGACATCAGCAAGTGAGTTCAT AA -5'	4655	68 °C
16S	5'- AGAGTTTGATCMTGGCTCAG -3' 3'- TACGGYTACCTTGTTACGACTT -5'	~1500	50 °C

3.9 Identification of lncRNAs

3.9.1 Systematic computational identification of lncRNAs

RNA-seq reads from the SRA at NCBI under the accession number SRP064341 were retrieved from a separate study on the *M. javanica* transcriptome performed concurrently by a colleague (Aini et al., 2016) during the study period. It is noted that RNA-seq data for *M. pentadactyla* was not available for comparison. Using the pangolin genome and annotations from the MAKER pipeline, reference-based assemblies were performed for each of the eight tissue samples with Tophat 2.0.8 (Kim et al., 2013) and Cufflinks 2.2.1 (Roberts, Pimentel, Trapnell, & Pachter, 2011; Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011; Trapnell et al., 2010, 2013). TopHat was used to map RNA-seq reads onto the pangolin genome using the '-G' option to include the annotation file from MAKER. The resulting BAM mapping file were filtered to only include the mapped reads and properly paired reads for assembly. The transcriptome was assembled using the MAKER annotation as reference in a Reference Annotation Based Transcript (RABT) assembly approach using Cufflinks 2.2.1. Cuffmerge was used to merge the assemblies from

different tissues. The resulting transcripts from the assemblies were subjected to a filtering pipeline as follows:

1. Transcript length: Multi-exonic lengths that were greater than 200nt were retained.
2. Known annotation: Cuffcompare from the Cufflink package was used to compare the annotation and resulting categories 'j', 'i', 'o', 'u' and 'x' were kept as possible non-coding transcripts.
3. Long ORF: Transcripts with long ORF of more than 300nt were excluded.
4. Exon Overlaps: Putative non-coding RNA was filtered if it intersected any known exon sequence of protein-coding loci in the reference annotation.
5. Coding potential: Coding potential of was predicted using PLEK based on an improved *k*-mer scheme (A. Li, Zhang, & Zhou, 2014). Besides that, CPC and CNCI was also used to assess transcript coding potential thereafter (Kong et al., 2007; Sun et al., 2013). Predicted non-coding transcripts was retained.
6. Protein domains: The resulting transcripts were translated to all three forward frames and NCBI CD-search was used to search each amino acid sequence against the Pfam database for known protein motifs with an e-value of 0.01 as threshold (Finn et al., 2015; Marchler-Bauer et al., 2002). Returned hits were excluded from the transcript list.

The final list of transcript was catalogued as high-confidence lncRNA transcripts of the *M. javanica* transcriptome. The computational pipeline is summarized in Figure 3.1.

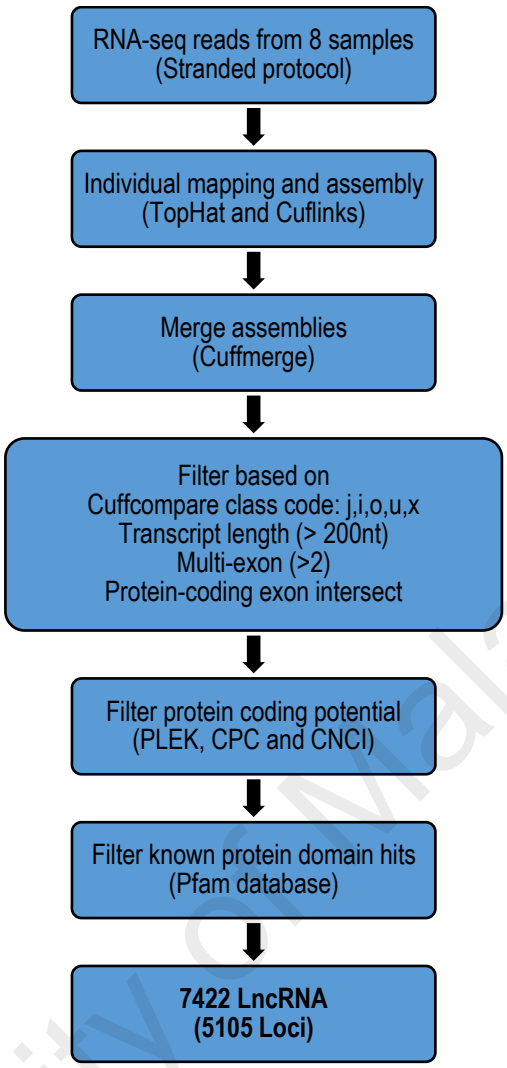

Figure 3.1 A systematic computational pipeline to detect and catalogue high confidence lncRNAs in *M. javanica*.

3.9.2 Classification of LncRNA by localisation

The set of lncRNAs found were classified by its localisation into antisense exonic, intronic and intergenic (lincRNA) in relation to known protein-coding loci. LncRNAs that are non-intersecting any known annotated loci are classified as lincRNA (intergenic LncRNA). LincRNA class forms the majority of the identified catalogue of lncRNAs. The remaining lncRNA by intersecting the respective loci were categorised into intronic or exonic. And, these lncRNA may be transcribed either in the sense or antisense direction. While the sense exonic intersecting transcripts were removed as mentioned in the previous section, only antisense exonic lncRNAs were considered in the exonic class.

3.9.3 Identification of repeats in lncRNA

Repeat content was identified using RepeatMasker (Smit et al., 1996) and using the Carnivora library from Repbase (Jurka et al., 2005) for each three subclasses of the lncRNA set. For comparison, the protein-coding mRNA was also searched with the same parameters.

3.9.4 Expression pattern of lncRNA

Differential expression of transcripts was identified for the eight tissue-type using Cufflinks against the merged assemblies of multiple tissues. The resulting expression unit in fragment per kilobase of transcripts per million (FPKM) mapped reads were normalised using a standard measure of $\log_{10}$ FPKM +1. The lncRNA transcript id was provided to the CumeRbund package (Trapnell et al., 2012) of R statistical package (R Core Team, 2014) which allowed the visualization of the differential expression heat map that was plotted based on the calculated $\log_{10}$ FPKM + 1 values.

CHAPTER 4: GENOME SEQUENCING AND ASSEMBLY

4.1 DNA extraction for genome sequencing

For the purpose of NGS, high molecular weight (HMW) and DNA of sufficient purity is required. This section describes the results obtained from DNA extraction from four sources – the blood, cerebrum, cerebellum, liver tissue of *M. javanica*.

4.1.1 Irregularities in DNA extraction from blood

One of the non-invasive ways of obtaining DNA from endangered and elusive species may be from hair bristles and animal droppings (Russello, Waterhouse, Etter, & Johnson, 2015). However, such methods have many challenges for NGS purposes. Specifically, the resulting DNA is often of low quality and low purity, and may lack the required integrity required for NGS. Here, attempts to extract DNA in a less invasive way from the blood of several male and female *M. javanica* were performed but was fraught with challenges.

Agarose gel electrophoresis of the DNA extracted from the pangolin blood with Qiagen Genomic Tips from 20/G DNA extraction kits showed peculiar banding patterns that are not commonly observed in other mammals (Figure 4.1). Such DNA fragmentation could be hallmarks of cellular apoptotic activity and commonly observed in cancer and sepsis (Almendro et al., 2003; Gavrieli, Sherman, & Ben-Sasson, 1992; Jahr et al., 2001). However, we need intact or high molecular weight DNA for whole-genome sequencing, therefore I decided not to use blood DNA for sequencing and instead extracted DNA from other pangolin tissues such as liver, cerebrum and cerebellum.

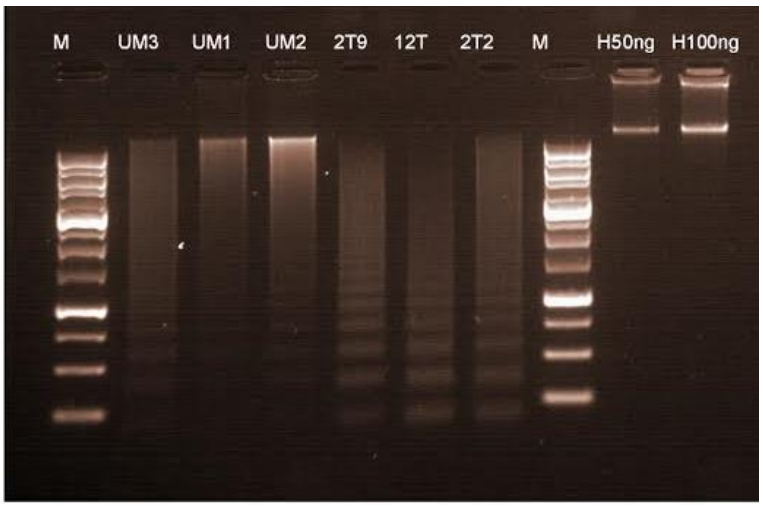

Figure 4.1 Agarose gel electrophoresis of DNA extracted from six blood samples collected from individual pangolin.
UM3, UM1 and UM2 are female pangolins while 2T9,12T, and 2T2 are male pangolins. Surprisingly, all samples showed DNA fragmentation patterns. M – DNA ladder; H50ng and H100ng – human blood.

4.1.2 DNA extraction from tissue samples for genome sequencing

Due to the fragmented DNA in blood DNA extraction procedure, DNA extraction was performed using liver, cerebrum and cerebellum tissue. The genomic DNA extracted from pangolin liver exhibited concentrations of more than 300 ng/ul and DNA quantitation exhibit A260/A280 ratios that range from 1.96 to 2.05 (Table 4.1). Mass and concentration requirements for NGS as per BGI (Hong Kong) guidelines were a minimum of 148 μg and 30 ng/μl for the eight short-insert libraries and four long mate-pair libraries planned. Corresponding gel electrophoresis of these DNA samples showed intact HMW bands that were suitable for NGS (Figure 4.2 and Figure 4.3). Since the extracted liver DNA had met the minimum requirements for NGS, I proceeded to use samples with ID GTNgs1, GTNgs2, BrNGSX1, and BrNGSY1, for whole-genome sequencing (Table 4.1).

Table 4.1 Concentration and purity of Qiagen Genomic Tip 20/G DNA extraction from pangolin liver, cerebrum and cerebellum measured by NanoDrop 2000 (Thermo Scientific).

Sample ID	Tissue	DNA Concentration (ng/ul)	Absorbance ratio (A260/A280)
GTNgs1	Liver	535.6	1.96
GTNgs2	Liver	383.9	1.97
GTNgs3	Liver	512.3	1.95
GTNgs4	Liver	351.7	1.99
GTNgs5	Liver	353.6	1.98
GTNgs6	Liver	552.7	1.96
BrNGSX1	Cerebrum	165.8	1.92
BrNGSX2	Cerebrum	172.3	1.92
BrNGSX3	Cerebrum	165.4	1.92
BrNGSY1	Cerebellum	267.0	1.88
BrNGSY2	Cerebellum	342.0	1.88
BrNGSY3	Cerebellum	398.3	1.87
BrNGSY4	Cerebellum	303.3	1.89

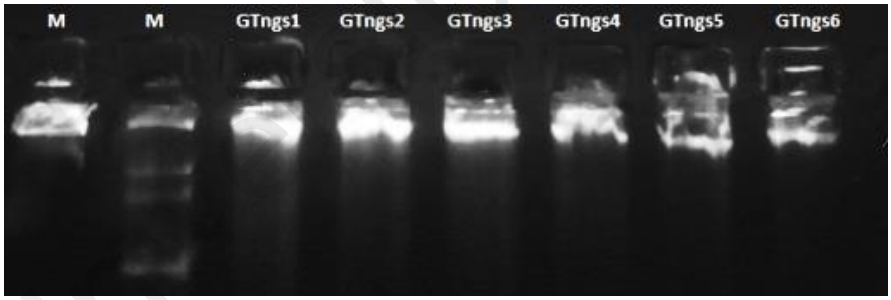

Figure 4.2 Gel electrophoresis of DNA extracts from liver using Qiagen Genomic tip 20/G.

Multiple batches of genomic DNA were visualized for integrity. M are lambda phage ladder and 2kb ladder respectively. Resulting bands showed all the samples had intact high molecular weight DNA, indicating high quality extracted DNA.

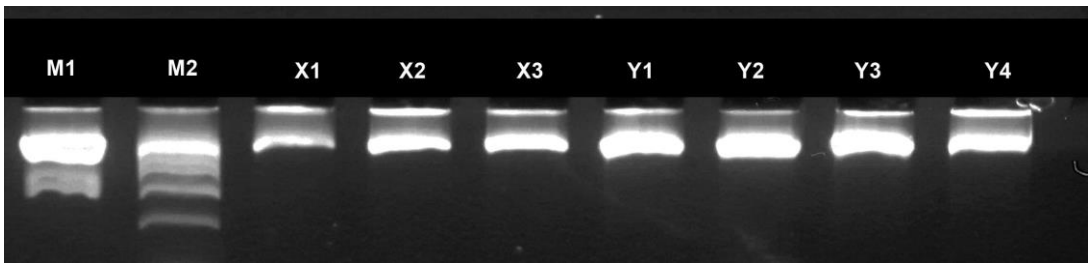

Figure 4.3 Gel electrophoresis of DNA extracts from cerebrum and cerebellum using Qiagen Genomic tip 20/G.
DNA from lanes name prefixed with X were extracted from cerebrum while Y were from cerebellum. M1 and M2 are lambda phage ladder and 2kb ladder respectively. Resulting bands showed all the samples had intact high molecular weight DNA, indicating high quality extracted DNA. *BrNGS should be prepended to match tag names as in Table 4.1.

4.2 Genome assembly from NGS raw data of *M. javanica*

The millions of sequences obtained from NGS in the form of FASTQ file must be assembled into the final genome. The subsequent sections outline the required natural steps that were used leading to a genome assembly.

4.2.1 Quality-check and pre-processing of sequencing raw data

The raw data that were obtained from the sequencing centre (BGI) were subjected to quality filtering and error correction as part of the pre-processing step. Each of the short read library was filtered with a Phred score of 20 (0.01% probability of error) for further downstream processing using PRINSEQ Lite (Schmieder & Edwards, 2011) and inspected further (Appendix 1) with the FASTQC program (Andrews et al., 2014). Each row in Table 4.2 represents the products of sequencing performed on different lanes of the flow cell after quality filtering. The resulting FASTQ files contains a particular number of reads which contributes to the overall sequencing depth of the sequencing project. The total coverage of NGS sequencing in that were obtained is 145-fold which was calculated based on the estimated genome size of 2.5 Gbp. The physical coverage that the read spans based on the library size for the *M. javanica* sequencing project is at an estimated 1192-fold. The depth of coverage used in this study was similar or more than

other comparable mammalian genome sequencing projects and deemed suitable for a draft genome assembly.

Table 4.2 NGS library reads statistics of *M. javanica*

Number of reads for each library and the estimated sequencing and physical coverage are shown. The sequencing coverage was estimated based on the predicted genome size of *M. javanica* (2.5Gbp) and *M. pentadactyla* (2.7Gbp) by *k*-mer analyses. Sequencing coverage the number of fold (X) of basepairs sequenced in relation to genome size. The physical coverage describe the coverage based on the constructed library length's base pair in relation to genome size.

Library	Tissue	Library Type	Total Paired-End (PE) Reads	Sequencing Coverage	Physical Coverage
M. javanica					
PE 180bp	Liver	Paired End	189,915,801	15.19	13.6
PE 180bp	Liver	Paired End	212,107,889	16.97	15.2
PE 500bp	Cerebrum	Paired End	160,493,968	12.84	32.0
PE 500bp	Cerebrum	Paired End	176,434,714	14.11	35.2
PE 500bp	Liver	Paired End	165,713,431	13.26	33.1
PE 800bp	Cerebellum	Paired End	115,914,999	9.27	37.1
PE 800bp	Cerebellum	Paired End	130,704,648	10.46	41.8
PE 800bp	Liver	Paired End	129,334,465	10.35	41.3
MP 2000	Liver	Mate-pair	115,649,108	9.25	92.5
MP 5000	Liver	Mate-pair	115,639,514	9.25	231.2
MP 5000	Liver	Mate-pair	154,755,684	12.38	309.5
MP 5000	Liver	Mate-pair	154,736,404	12.38	309.5
			Total:	146.07	1192.5

4.2.2 *M. pentadactyla* NGS raw reads and assembly data

The NGS data in the form of FASTQ files and genome sequence of *M. pentadactyla* was provided by our collaborators from University of Washington and the read and genome assembly statistics are shown in Table 4.2 and Table 4.3. This will serve as background information for the comparative genomics analysis in subsequent sections.

Table 4.3 NGS library reads statistics of *M. pentadactyla*
Number of reads for each library and the estimated sequencing coverage are shown. The sequencing coverage was estimated based on the predicted genome size of *M. javanica* (2.5Gbp) and *M. pentadactyla* (2.7Gbp) by *k*-mer analyses.

Library	Library Type	Total Paired-End (PE) Reads	Sequencing Coverage	Physical Coverage
M. pentadactyla				
PE 206	Paired End	437,447,710	32.40	33.41
PE 356	Paired End	296,308,797	21.94	39.11
MP 3000	Mate-pair	25,634,681	1.90	28.51
MP 8000	Mate-pair	1,172,455	0.08	3.48
			Total: 56.32	104.51

Table 4.4 Summary genome assembly statistics for *M. pentadactyla*.
The genome was assembled using SOAPdenovo

Assembly	Contigs	Scaffold
# contigs/scaffold (>= 0 bp)	230,930	87,621
# contigs/scaffold (>= 1000 bp)		33,682
Total length (>= 0 bp)	1,999,057,008	2,205,289,822
Largest contig/scaffold	292,755	1,402,852
N50	28,718	157,892
# N's per 100 kbp	0	9407.54

4.2.3 Read-based characterisation

Prior to a complete assembly of the genome, it is useful to identify and assess the inherent quality of the NGS data. By performing read characterisation, several important information such as variant and repeat frequencies, read coverage and GC-bias information can be garnered. Here, these features in the *k*-de Bruijn graph were inferred prior to genome assembly using read characterisation for both *M. javanica* and *M. pentadactyla* with a few other species of bird (*Melopsittacus undulates*), fish (*Maylandia zebra*), oyster (*Crassostrea gigas*), and snake (*Boa constrictor constrictor*) as reference.

Variant branching frequency of *M. pentadactyla* appear to be lowest among the vertebrate species while *M. javanica* shows high levels but lower than fish and oyster (Figure 4.4 a). Thus, the *M. javanica* genome is expected to be highly heterozygous like the bird genome and is expected to have higher heterozygosity compared to the *M. pentadactyla* which possesses lower variant branches. Like the oyster genome, high number of repeats content is to be expected in both pangolin genomes (Figure 4.4 b). The 51-kmer histogram (Figure 4.4 c) followed a bimodal curve for both *M. javanica* and snake data while the *M. pentadactyla* curve appears unimodal. The read coverage of the raw data can thus be inferred and inform whether the DNA was sequenced at sufficient depth. The second peak observed in Figure 4.4c is the true genomic *k*-mers of which when abundant, aids to better genome assembly by differentiating true genomic *k*-mers against error *k*-mers. The GC-bias plots which show secondary clumps often denotes the presence of contaminants as in the case of *M. javanica* (Figure 4.4e). Genome assemblers are limited in ability to assemble genomes into highly contiguous scaffolds (high N50) when the genome is sequenced at a low coverage, highly heterozygous, or contain many repetitive sequences. Although our genome was sequenced at a high coverage of approximately 145-fold, the

genome assembly was foreseen to be challenging due to high variant branches and repeat branches.

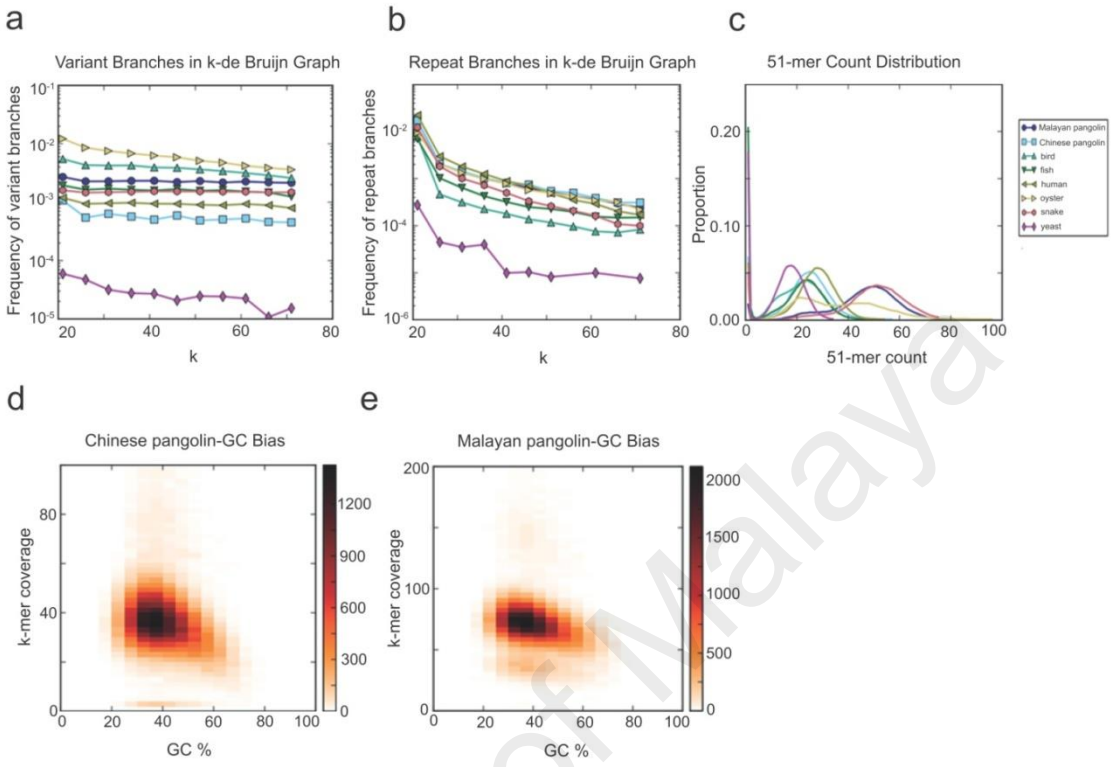

Figure 4.4 NGS reads-based characterisation of pangolin genome.
(a) Frequency of variants branches in the *k*-de Bruijn graph (b) Frequency of repeats branches in the *k*-de Bruijn graph. (c) 51 *k*-mer distribution to infer the coverage of the reads. (d) GC content plot of the *M. pentadactyla* reads. (e) GC content plot of the *M. javanica* reads.

4.2.3.1 Genome size estimation

There was no available report on the genome size of *M. javanica* and *M. pentadactyla* either in literature nor the Animal Genome Size database (Gregory et al., 2007). Previously, it was shown that genome size predictions derived from read characterisation closely represent actual values derived from wet-lab experiments (Simpson, 2013). Based on *k*-mer counting, the predicted genome size of the *M. javanica* and *M. pentadactyla* is predicted to be approximately 2.5Gbp and 2.7Gbp, respectively (Figure 4.5 and Table 4.5).

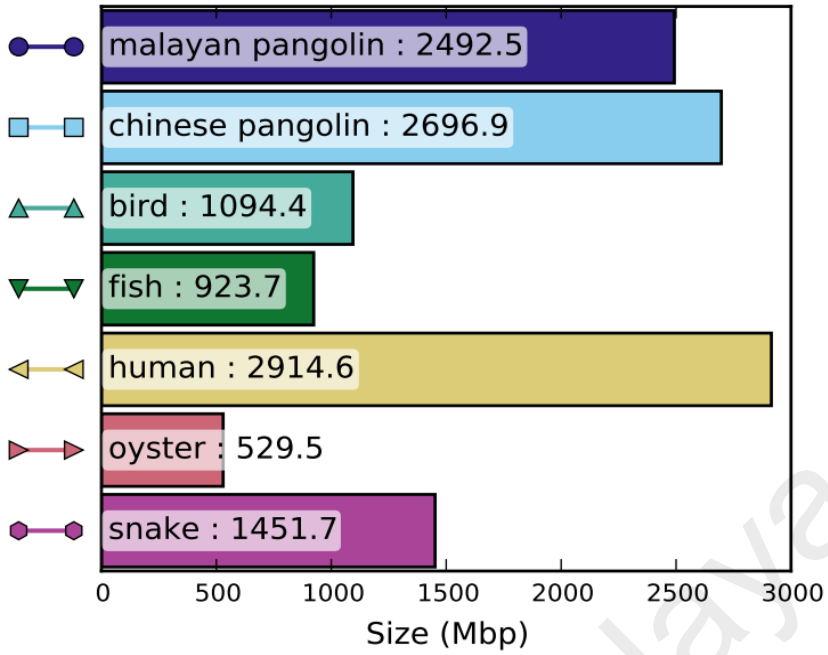

Figure 4.5 Genome size prediction of the Malayan (*M. javanica*) and Chinese pangolins (*M. pentadactyla*) compared to other organisms using 31-mer counting
The datasets of other organisms used in this analysis were from previous study (Bradnam et al. 2013). Bird=*Melopsittacus undulatus*; Fish=*Maylandia zebra*; Oyster=*Crassostrea gigas*; and *Snake=Boa constrictor constrictor*.

Table 4.5 Genome size estimation using SGA preqc module.

	M. javanica	*M. pentadactyla*
Estimated genome size	2,492,544,425 bp	2,696,930,760 bp

4.2.4 Genome assembly

The genome of *M. javanica* were performed using three assemblers: CLC Assembly Cell (version 4.10, CLC bio, Aarhus, Denmark), SGA (Simpson & Durbin, 2010) and SOAPdenovo2 (R. Luo et al., 2012). The assemblers were tested by parameter exploration to obtain the optimal assembly through its crucial parameter such as *k*-mer or overlap value (specific to SGA only). Optimal *k*-mer values for *M. javanica* dataset were 43 and 33 for SOAPdenovo2 and CLC assembly cell, respectively. As for SGA, the overlap value of 81 was optimal (Figure 4.6).

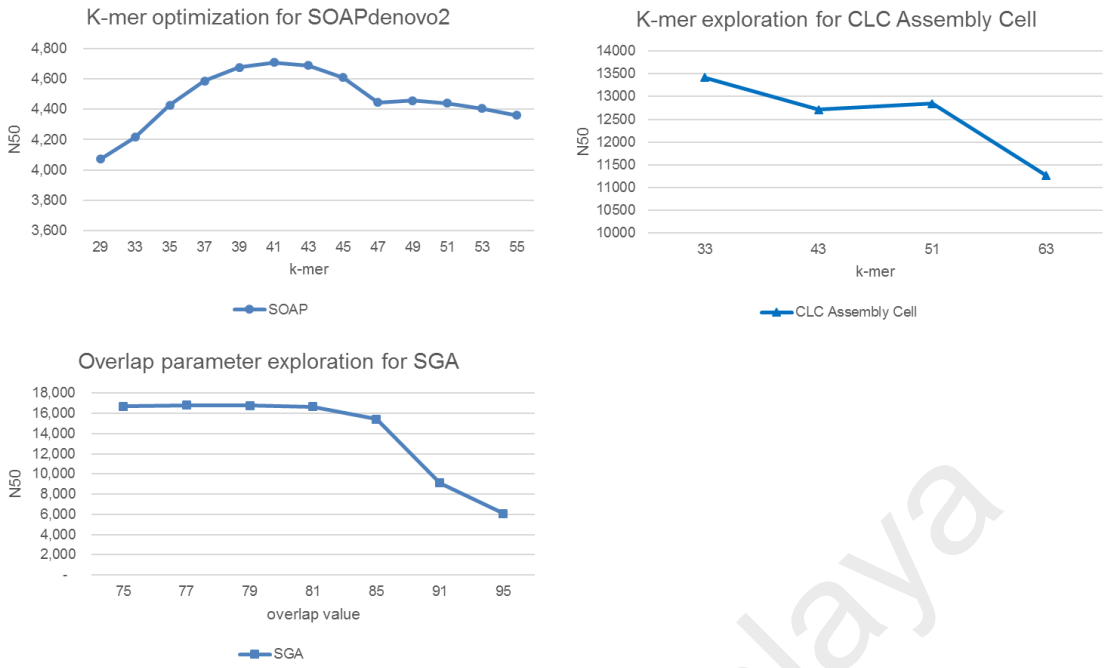

Figure 4.6 Parameter exploration for various assemblers.
Multiple rounds of assembly were performed with *k*-mer values varied for SOAPdenovo2 (29-55) and CLC Assembly Cell (33-63). In SGA, overlap value was tested (75-95).

Each optimal assembly obtained from the three assemblers (Figure 4.6) were compared for precision using FRCbam (Narzisi & Mishra, 2011). The FRCbam program calculates various types of suspicious features using the alignment file of reads against each contig in the assembly and outputs values that can be used to construct a Feature Response Curve (FRCurve) (Vezzi et al., 2012). Typically, the genome coverage should increase sharply as an indication of a good assembly i.e. the slope of FRCurve should be steep. Based on the plot (Figure 4.7), the steepest curve of SGA showed the highest genome coverage to feature value, therefore I selected SGA as the most optimal assembly in our tests for other downstream analyses. Here, the SGA genome assembly level was still at the contig stage that requires further scaffolding and gap-closing steps.

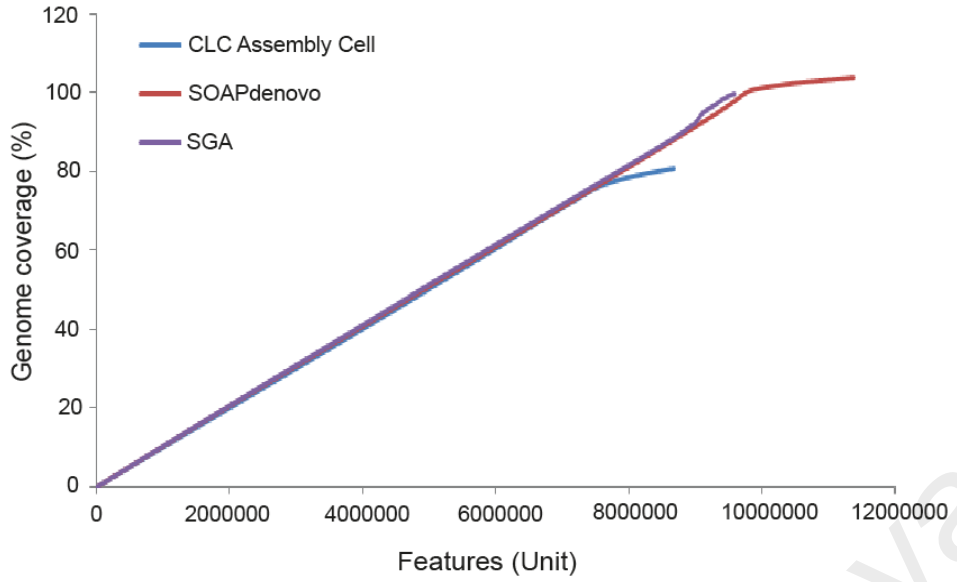

Figure 4.7 FRCurve of three assemblies of *M. javanica.*

4.2.5 Final draft genome assembly

The final assembly was produced based on SGA-generated contigs after scaffolding with available mate-pair libraries to place the assembled contigs into larger scaffolds. The scaffolding performance of *M. javanica* using SOAPdenovo2 scaffolder using SOAPdenovo2 contigs and SGA contigs are shown (Figure 4.8) and scaffolding of panda contigs from a previous study is given as a reference. The final draft assembly was gap-closed and any scaffolds smaller than 1000bp were discarded. The final draft assembly was 2,549,959,554 bp in size and hence forth is referred to as the pangolin assembly or *M. javanica* assembly (Table 4.6).

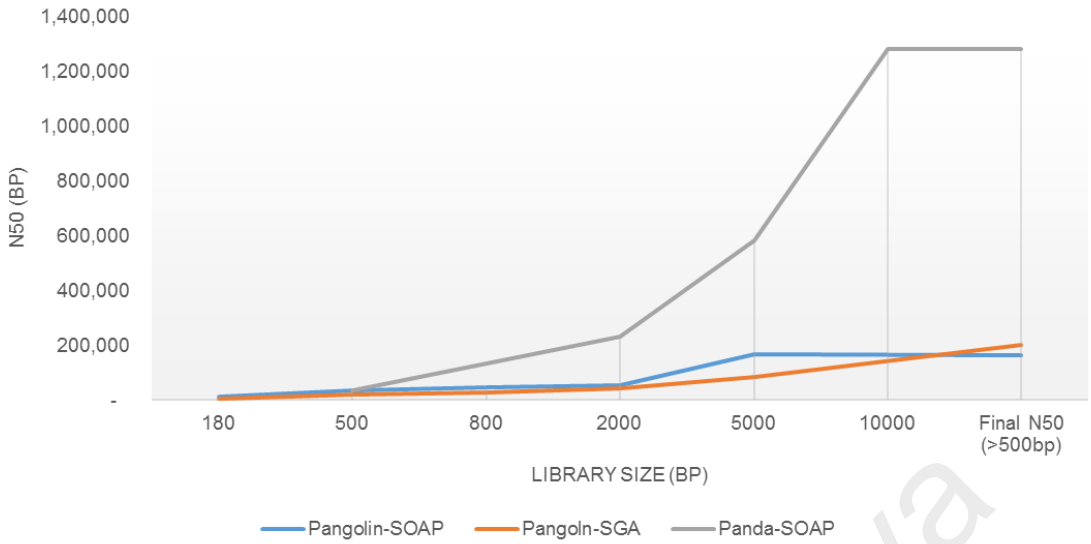

Figure 4.8 Increases in scaffold N50 by library utilisation during scaffolding by SOAPdenovo2.
This figure depicts the increasing trend of N50 statistics when increasing sizes of paired-end library was used to perform scaffolding. As a comparison, the increases of N50 in panda genome scaffolding was noted.

Table 4.6 Summary statistics for the assembled genome of *M. javanica*.

	Paired-end insert size	Estimated Coverage (X)	N50 (bp)	Total length (bp)
Initial contig (>1k)	180bp, 500bp, 800bp	102.45	17,568	2,047,445,145
Final scaffold (>1k)	180bp, 500bp, 800bp, 2000bp, and 5000bp	145.71	204,525	2,549,959,554
Final contig (after gap-closing)			18,812	2,108,780,390

4.2.6 Quality evaluation of pangolin genome assemblies

4.2.6.1 Transcript mapping

Both pangolin assemblies were mapped using the set of 89,754 consensus transcripts (as well as the set of 1,035,201 transcripts generated by Trinity alone (Grabherr et al., 2011) using GMAP software (T. D. Wu & Watanabe, 2005). The summary of the percentage of mapped consensus UniGenes and the Trinity-generated transcripts were calculated (Table 4.7). In general, at least 93% of the transcripts were mapped to both *M. javanica* and *M. pentadactyla* assemblies for both sets of pangolin transcripts indicating the high quality of our assemblies for genome annotation.

Table 4.7 Transcript mapping of Unigenes and Trinity-assembled transcripts using GMAP.

89,754 Consensus Unigenes	***M. javanica***	***M. pentadactyla***
Unmapped transcripts	282 (0.31%)	4654 (5.19 %)
High quality mapped transcripts (alignments with 92% identity and 80% coverage)	75321 (83.91%)	58125 (64.76 %)
Mapped transcripts	14151 (15.76%)	26975 (30.05%)
1,035,201 Trinity-generated Transcripts	***M. javanica***	***M. pentadactyla***
Unmapped transcripts	8419 (00.81%)	58326 (5.63 %)
High quality mapped transcripts (alignments with 92% identity and 80% coverage)	907603 (87.67%)	712245 (68.80 %)
Mapped transcripts	119179 (11.51%)	264630 (25.56%)

Due to the fact that the RNA transcripts were assembled independently, high mapping rate of the RNA transcripts onto the pangolin genome assemblies further support the high quality of the assemblies and reinforce the suitability of the draft genomes for other downstream analysis.

4.2.6.2 CEGMA analysis

CEGMA is a computational method that relies on a defined set of ultra-conserved eukaryotic protein families for building a highly reliable set of gene annotations (Parra et al., 2007). The CEGMA analyses indicated 91% ultra-conserved eukaryotic genes present in the *M. javanica* assembly and 58% were considered complete genes. For *M. pentadactyla* assembly, a total of 88% of ultra-conserved eukaryotic genes were present in this genome, 55% of which were considered complete genes. The gene space completeness statistics showed that the pangolin genome is a good candidate for genome

annotation and subsequent analysis as it has high gene space completeness level similar to other eukaryotic genome projects.

4.3 Presence of bacteria in pangolin samples

After genome assembly, bacterial contamination screening was performed after finding a distinct island in the GC heatmap plot (Figure 4.4e) and found bacterial sequences in the pangolin genome assembly. To further investigate this, we screened the tissue-specific genome sequence from three different tissues (cerebrum, cerebellum and liver) that were independently assembled. During the bacterial contamination screening using BLASTN (Altschul et al., 1997) against the bacterial nucleotide sequence database (with 97% sequence identity and coverage), a relatively large number of bacterial sequences were detected in the pangolin assemblies of cerebrum and cerebellum, but not in the liver (Table 4.12). The contig sequences of the assembled cerebral genome, resulted in 5890 BLAST hits with known bacterial sequences, with the all those hits (100%) having best matches to *Burkholderia fungorum* ATCC BAA-463 (total genomic length = 6,324,546 bps). In the assembled cerebellum genome, contig sequences had 3004 hits to known bacteria. Of those hits, majority (99.33%) were likely from *B. fungorum* (962,651 bps), indicating that the pangolin cerebrum and cerebellum tissues were likely colonised by *B. fungorum*, although these tissues should be sterile.

Table 4.8: Detection of bacterial sequences in the pangolin tissue-specific genomes. 6635 contigs in cerebrum specific genome data with 6,818,896 bps of *B. fungorum* ATCC BAA-463 bacteria hit found, whereas 3533 contigs in cerebellum specific genome data with 1,109,334 bps observed. No significant matches to *B. fungorum* ATCC BAA-463 observed in the liver assembly. (90% sequence identity and completeness)

Threshold	**90% identity 90% coverage**		**97% identity 97% coverage**	
Tissue-specific genome	**Cerebrum**	**Cerebellum**	**Cerebrum**	**Cerebellum**
Total contigs in pangolin with bacteria homologs	6730	3533	5890	3004
*Bf *ATCC BAA-463*	6635 (98.58%)	3452 (97.70%)	5890 (100%)	2984(99.33%)
# of base pairs	6,818,896 bp	1,109,334 bp	6,324,546 bp	962,651 bp
Other bacterial species	95	81	0	9

*Bf - *Burkholderia fungorum*

4.3.1 *B. fungorum* screening across different adult tissues

To examine whether *B. fungorum* is also present in other pangolin tissues, PCR assays were performed to amplify target regions using a set of primers specific to *B. fungorum* in nine different tissues (cerebrum, cerebellum, liver, blood, kidney, thymus, spleen, lung, and heart). To include controls and capture different scenario, three pairs of *Burkholderia*-specific primers were designed (to detect whether other *Burkholderia* species are present if *B. fungorum* is not present) and a pair of universal bacterial 16S gene primers (to detect whether other bacterial species are present in the absence of *Burkholderia* species).

PCR assays showed clear bands in the cerebrum, cerebellum, blood and lung, but not in other tissues such as liver and kidney (Figure 4.16), indicating the presence of *B. fungorum* in the four tissues. Interestingly, a significant band in the blood sample suggest that the female *M. javanica* used in this study might have developed septicaemia.

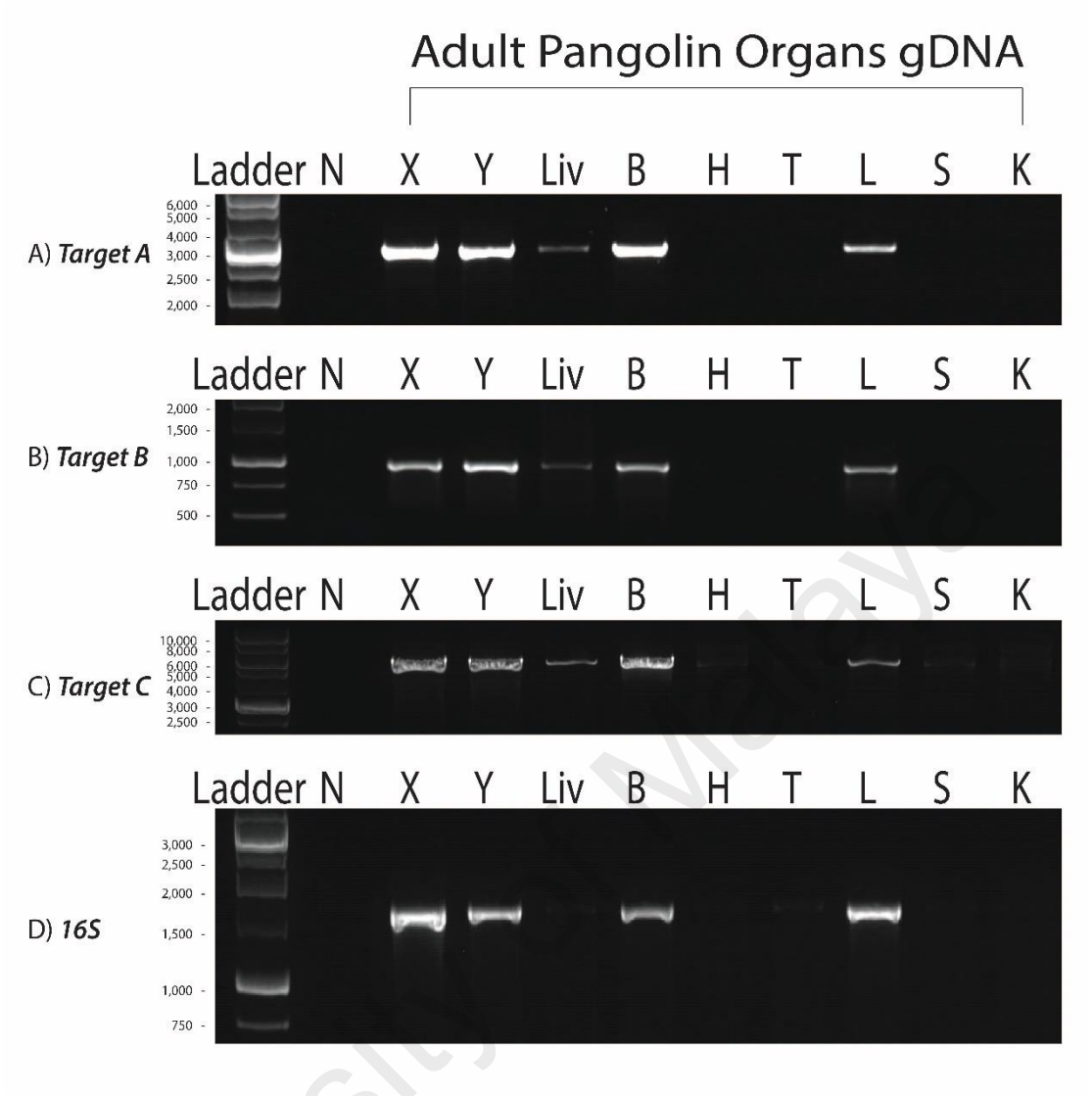

Figure 4.9 PCR amplification of *B. fungorum* targets across nine tissues in the same individual pangolin

(Target A - transposase, Target B - OI25_7129 hypothetical protein, Target C – DNA polymerase; N- Negative control, X- Cerebrum, Y- Cerebellum, Liv- Liver, B- Blood, H- Heart, T- Thymus, L- Lung, S- Spleen and K- Kidney) (A) PCR result for target A. Positive bands present at approximate 3kb. Positive bands found in X, Y, Liv, B, and L. (B) PCR results for target B. Positive bands observed at approximate 1kb. Positive bands found in X, Y, Liv, B and L. (C) PCR results for target C. Positive bands observed at approximate 5kb. Positive bands found in X, Y, Liv, B and L. (D) PCR results of bacteria 16S. Positive bands observed at approximate 1.5kb. Positive bands found in X, Y, B and L.

4.3.2 Presence of *B. fungorum* in other pangolins

To examine whether the presence of *B. fungorum* in *M. javanica* organs is an isolated case, I also screened the blood of other individual pangolins using PCR assays with the same set of primers. The DNA derived from blood of seven adult pangolins (tags: UM1,

UM2, UM3, 26T, 2T9, 12T and 2T2) were used in screening for *B. fungorum* (Figure 4.17). Interestingly, of the seven pangolin blood samples, four samples were found to have clear and consistent positive bands for the set of primers used. This result also supports the view that the presence of *B. fungorum* in the pangolin (UM3) was not an isolated case. It should be noted that although significant bands in three pangolin blood samples were not detected, it does not mean the pangolins were not colonised by *B. fungorum*. The possibility that colonisation by *B. fungorum* in other tissues e.g. lung and brain, that is yet to spread to the blood could not be ruled out.

Figure 4.10 PCR assays using the blood of seven pangolin individuals
Using the three same targets as previously described to identify the infected pangolin. For all PCR, N is the negative control and UM3 is from the host pangolin used as positive control (genomic and PCR evidence). UM1, UM2, UM3, 26T, 2T9, 12T and 2T2 are amplifications from DNA extraction of individual pangolins. Results show positive bands in UM3, 2T9, 12T and 2T2, likely indicating the presence of Bf in these tissues.

4.4 Summary

In this chapter, major results from the pangolin genome sequencing, assembly and analysis were highlighted. I introduced results from DNA extraction and quality checking using agarose gel electrophoresis and Nanodrop quantitation of DNA content and purity. After DNA sequencing, the resulting sequencing raw reads statistics were

given and coverage were estimated. Several steps in genome assembly involving read characterisation, contig assembly, assembly comparison, scaffolding and quality evaluation was systematically presented. Finally, we share the unexpected finding of genomic screening which revealed the colonisation of *B. fungorum* in multiple tissue samples and in different individuals.

CHAPTER 5: GENOME ANNOTATION AND EVOLUTIONARY ANALYSIS

5.1 Genome structural and functional annotation

5.1.1 Repetitive sequences

The genome sequences of eukaryotes are pervasive with repetitive sequences also known as transposable elements. It is important to characterize and mask these repeats prior to gene annotation as its repetitive sequences are known to interfere gene-finding software.

5.1.1.1 Repeat annotation

There are various classes of repeat sequences such as LINEs, SINEs, LTR elements and DNA elements. Repeatmasker software is able to detect repetitive sequences based on homology to a repeat library. Generally, the major repeat classes of both pangolin species were similar (Table 5.1) although a greater amount of DNA was masked when the carnivore library was used in RepeatMasker. As much as 3-4% of LINE sequences were masked using carnivore library as opposed to using the mammal library.

Table 5.1 Proportion of common repeat elements in *M. javanica* and *M. pentadactyla* detected using different libraries

	M. javanica		***M. pentadactyla***	
Library	mammal	carnivora	mammal	carnivora
LINE	18.74%	21.80%	17.36%	21.72%
SINE	2.61%	2.63%	2.79%	2.82%
LTR Elements	6.05%	4.93%	5.65%	5.50%
DNA Elements	2.90%	2.76%	3.02%	3.13%

Overall, the distribution of the predicted repeat sequences of the pangolins and other mammals were highly similar, except the SINE family (Figure 5.1). The percentage of

the SINE family in pangolin genomes is the lowest (2.6 %) compared to other mammals (3.2 – 12.7%).

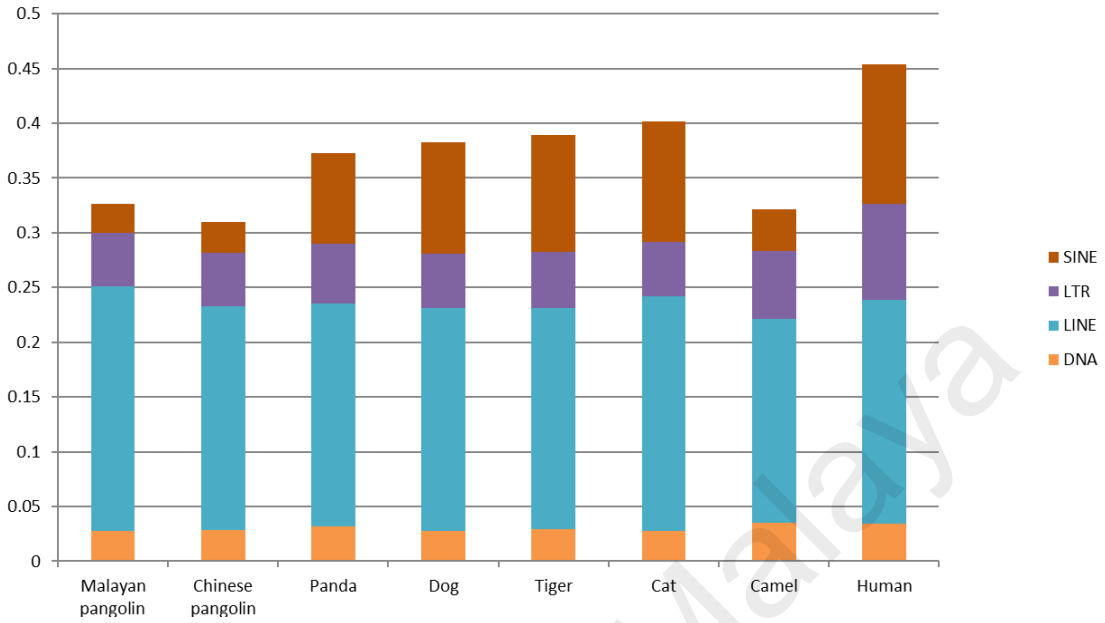

Figure 5.1 Repeat sequences present in genomes of closely related mammals

To investigate the possible reason on why the pangolins have significantly low proportion of SINE family, I further examined the distribution of SINE families. Interestingly, I found that pangolins lack of tRNA-SINEs has not been observed before in other mammalian genomes when other form of SINE such as MIR or Alu is present except in megabat (*Pteropus vampyrus*) (Figure 5.2). Nonetheless, there are reports of complete absence of the whole SINE class in non-mammalian species such as *Drosophila* (fruit fly) and *Galus galus* (chicken) (Kramerov & Vassetzky, 2011).

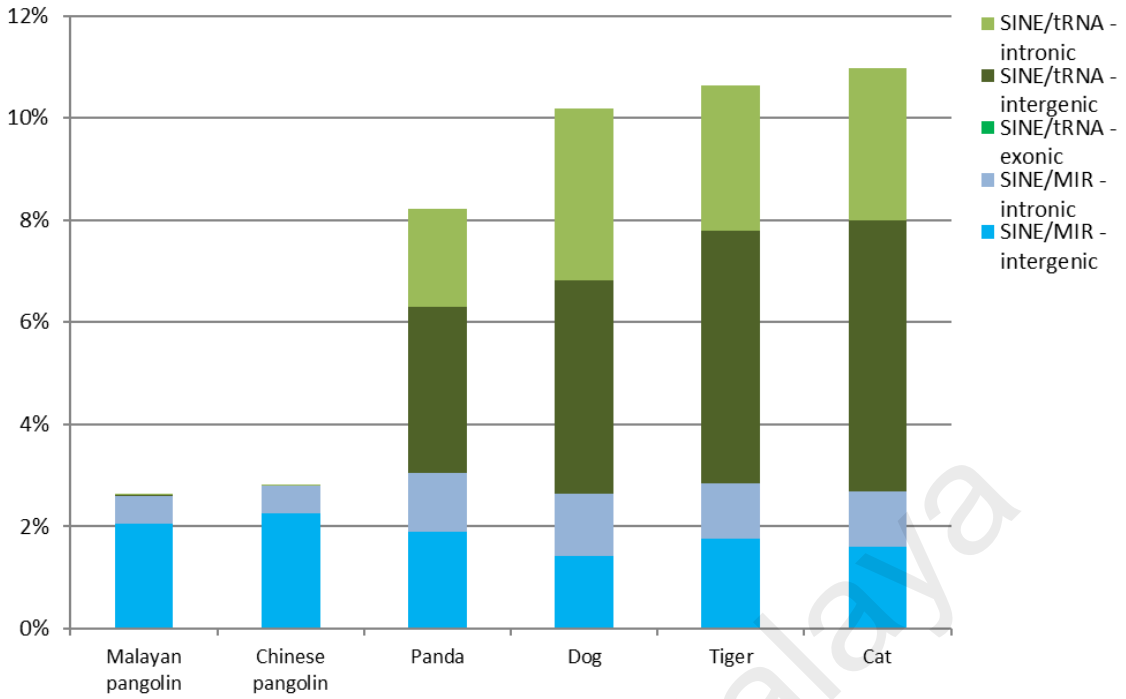

Figure 5.2 Distribution of SINE families across closely related Carnivoran species. Comparison of distribution of SINE families between Carnivora species were also shown by gene structure composition.

5.1.1.2 *De novo* repeat sequence identification

While the usual Repeatmasker method is suitable to search and annotate repeat sequences, however, in non-model organisms which may not yet have a robust repeat library, a *de novo* approach to find novel repeats may be useful. *De novo* repeat finding was performed in hope to recover any missed SINE sequences from the previous method (Table 5.2). However, the findings concurs with the above findings, that SINEs are proportionately low in the genome of pangolins compared to other repeat classes.

Table 5.2 Repeat sequence content in pangolins as per *de novo* repeat-finding method

	M. javanica	*M. pentadactyla*
LINE	19.19%	16.37%
SINE	1.41%	1.11%
LTR Elements	5.83%	5.07%
DNA Elements	1.86%	1.81%

5.1.1.3 Repeat composition

Once the gene annotation was known, the repeat composition by gene structure can be described as it is known to traverse across various gene boundaries. Here, a relative composition of repeat elements could be visualized among Carnivora species based on gene structure and intergenic space in Figure 5.3. In general, it is observed that pangolins have a lower composition of SINEs. Particularly in pangolins, intronic SINE composition is the lowest among the mammalian species investigated (Figure 5.3). Further inspection revealed that the low proportion of SINEs in pangolins is likely due to the lack of tRNA-derived SINEs when compared among closely related carnivore species (Figure 5.1).

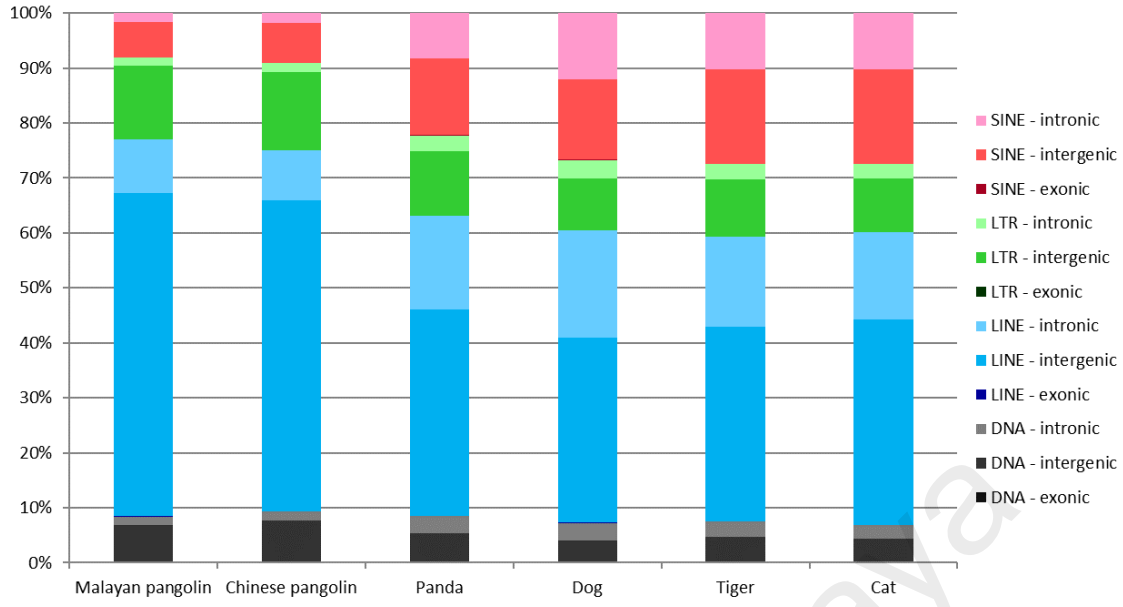

Figure 5.3 Repeat composition by gene structure among closely related Carnivoran species.
Identified repeat families that overlap intronic, intergenic, and exonic regions of the genome is shown with hues of red, green, blue and grey.

5.1.1.4 Repeat landscape

The repeat landscape depicts the transposable elements families that were propagated within *M. javanica* and *M. pentadactyla* along evolutionary timescales (shown by Kimura substitution levels). Both repeat landscapes showed a highly similar pattern of transposable elements propagation (Figure 5.4). The repeat activity of almost all classes show a decline (inactivation) at about 27 Kimura substitution (approximately 64 MYA). The period coincides with the split of the common ancestor of Carnivora and Pholidotans (Figure 5.6). It suggests that repeat activity was declining ever since the common ancestor of the pangolin appeared 64 MYA.

In the previous section, the genome size estimation was shown to differ greatly among *M. javanica* and *M. pentadactyla*. It was predicted that *M. pentadactlya* has larger genome size compared to *M. javanica*. It is possible that the genome expansion found in *M. pentadactyla* can be explained by duplication events, and repeat activity. However, our

data clearly showed that the repeat activity for both the genome appear closely similar with no significant pervasive propagation of repeat elements that occurred recently. Hence, we cannot rule out the possibility that duplication activity could be a responsible force in shaping the expansion of the genome in *M. pentadactyla* explaining the marked difference in the genome size estimates.

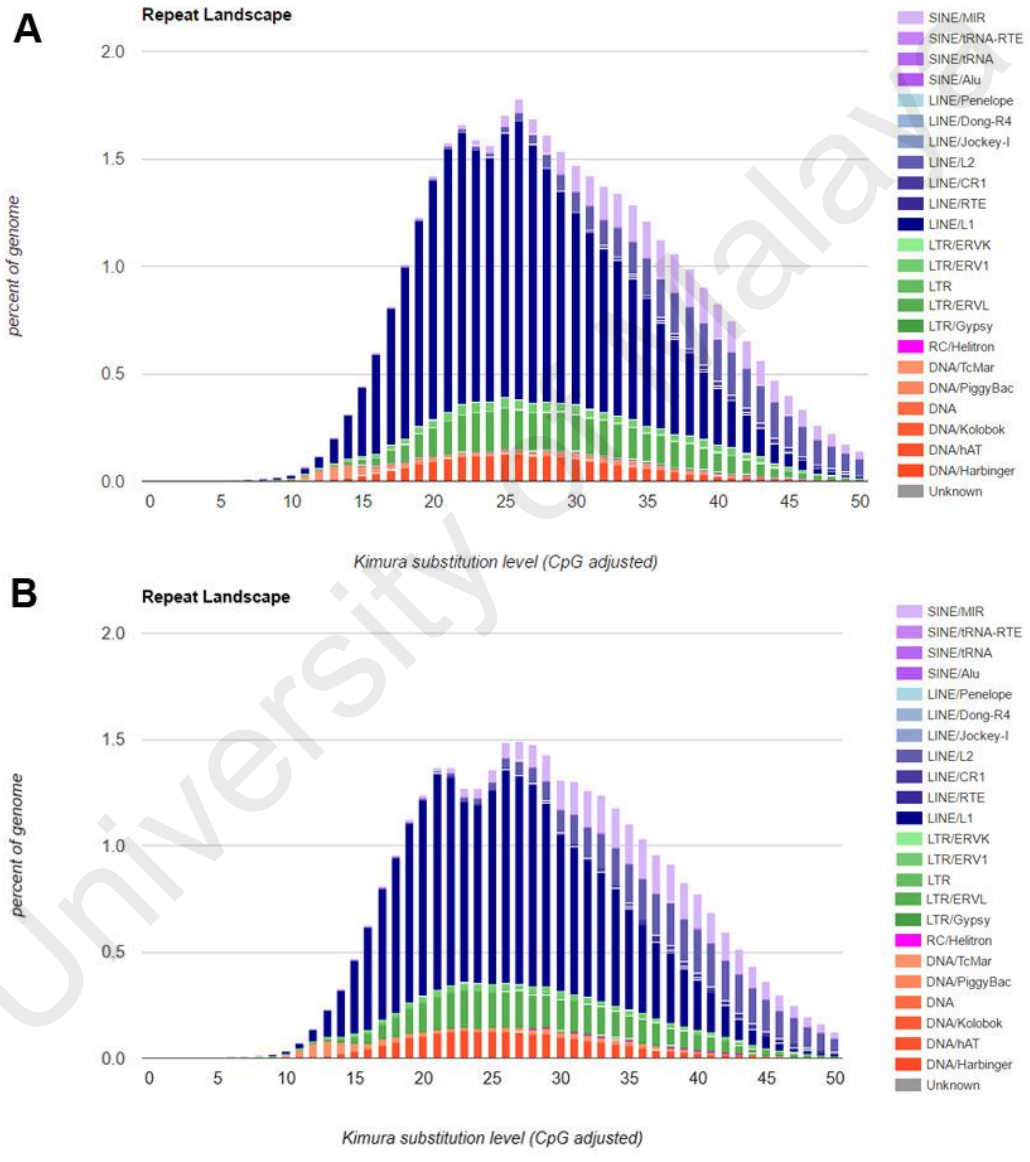

Figure 5.4 Repeat landscape of (A) *M. javanica* and (B) *M. pentadactyla*
Repeat landscape of pangolins describes the expansion and contraction of transposable element classes.

5.1.1.5 Segmental duplications

To examine the segmental duplications in the pangolin genomes, we identified it using a in-house developed pipeline from LAST, clasp and EMBOSS packages. Segmental duplications are commonly defined as sequences that are present in the genome that are at least 90% in sequence similarity with each other and are at least longer than 1000 bp. I identified a total of 21,843 putative duplicated fragments spanning 36.28 Mb (1.45%) in the *M. javanica* genome and 19,621 fragments spanning 44.80 Mb (1.66%) for *M. pentadactyla* (Table 5.3). It should be noted that the percentage of the genome coverage of segmental duplication coverage was calculated by assuming the genome size of *M. javanica* and *M. pentadactyla* are 2.5 Gb and 2.7 Gb, respectively. Generally, segmental duplication are known drivers of genome evolution particularly responsible for the emergence of new gene functions and chromosome instability (Bailey, Church, Ventura, Rocchi, & Eichler, 2004). When pervasive, it could cause the increase in genome size (genome expansion).

Table 5.3 Summary of segmental duplications in the pangolin genomes. Different cut-off sizes were used to detect the segmental duplications in the pangolin genomes.

SD Size	***M. javanica***			***M. pentadactyla***		
	Number of SD	Median	Genome Coverage (Mb)	Number	Median	Genome Coverage (Mb)
> 1KB	21,843	1,215	36.28	19,621	1,438	44.80
> 5KB	439	7,214	7.85	855	16,206	14.80
> 10KB	135	29,014	5.78	675	18,115	13.50
> 50KB	45	73081	3.78	0	0	0

*SD – Segmental duplication

5.1.2 Gene annotation

To identify putative genes in the pangolin genomes, I used the MAKER annotation pipeline based on several evidence sources, primarily, *ab initio* gene prediction, transcriptomic data from *M. javanica* and protein evidence from *Canis familiaris* reference genome and transcriptome. Using this approach, I identified a total of 23,446 and 20,298 genes in the *M. javanica* and *M. pentadactyla* genomes respectively (Table 5.4). The *M. javanica* genome has 170,236 exons, whereas the *M. pentadactyla* genome has 147,455 exons. The total exon length accounted for approximately 1.81% and 1.34% of the *M. javanica* and *M. pentadactyla* genomes, respectively. The resulting protein-coding genes identified were the product of evidence-based annotation using the non-redundant Refseq protein database. The number of genes found in pangolins were typical of mammalian genomes such as human (30,804), cat (19,493), dog (33,202), rat (27,490), and panda (19,343) (Hou & Lin, 2009; Hubbard et al., 2009).

Table 5.4 Gene structure annotation from MAKER annotation pipeline

	M. javanica	***M. pentadactyla***
Number of genes	23,446	20,298
Number of exons	170,236	147,455
Number of five_prime_UTR	10,925	8,535
Number of three_prime_UTR	10,919	8,817
Total exons length (bp)	45,284,578	36,081,291

Upon comparing its gene features across vertebrate annotations available on ENSEMBL, the mean intron length in pangolins were found to be the lowest among other mammals (Figure 5.5). Previous studies have associated intron length to flying and no-flying avian groups, suggesting a link between metabolism and intron length.

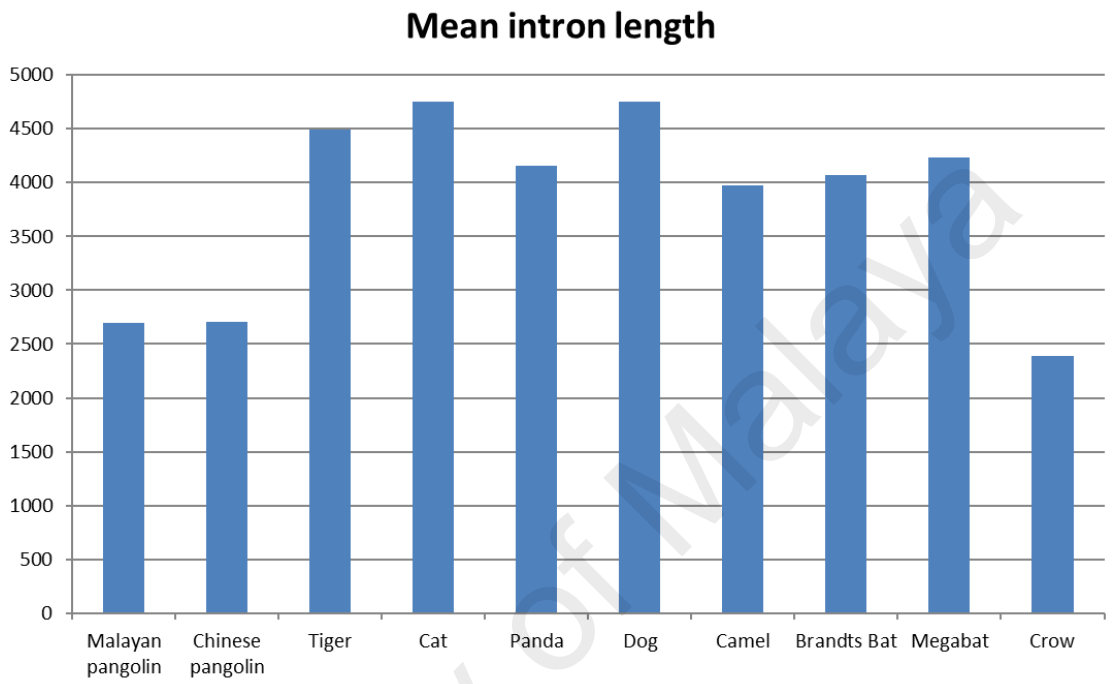

Figure 5.5 Mean intron length of vertebrate species.
Histogram shows the average intron length across different vertebrate species. Pangolins have the lowest intron lengths among mammals.

5.2 Evolutionary Analysis

The following sections describe the evolutionary analysis that were performed using the nuclear genome.

5.2.1 Speciation Time and Divergence Time

By using molecular clock and fossil records, the speciation and divergence times of the pangolin and their closely related relatives can be inferred. Our data showed that (Figure 5.6) pangolins diverged from their closest relatives, the Carnivora, around 56.8-67.1 millions years ago (MYA) (mean=61.9MYA). *M. javanica* and *M. pentadactyla* species diverged from each other approximately 4-17.3 MYA (mean=8.84MYA). Nucleotide

divergence between pangolins for 4-fold degenerate sites was calculated and found to be 0.42 % per million years. Based on this, substitution/mutation rate, 1.47 x 10^{-8} with 95% interval (3.25 x 10^{-8} -7.51 x 10^{-9} for minimum and maximum divergence time) were calculated.

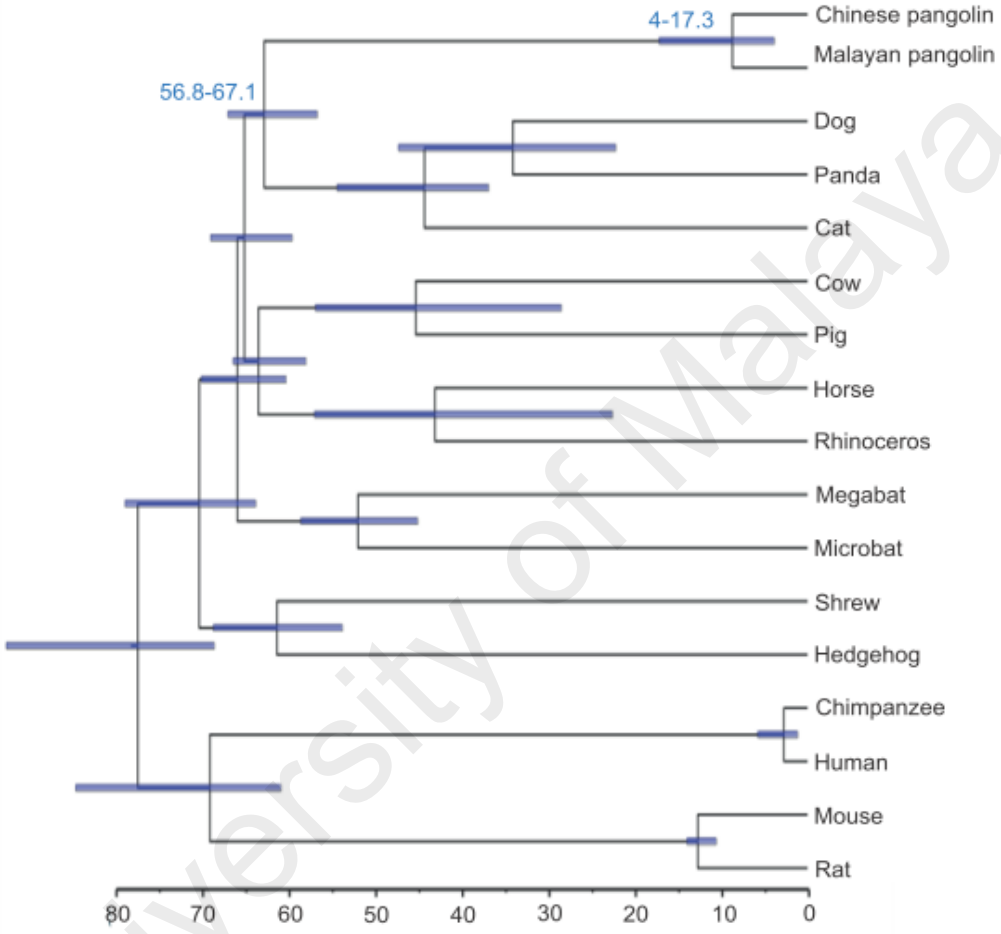

Figure 5.6 Estimation of speciation and divergence time.
Chronogram with 95% intervals of posterior divergence time (in millions of years) distribution of 17 mammalian species calibrated with fossil information. Posterior distributions of divergence times (blue numbers on the nodes) of Pholidota lineage and Chinese (*M. pentadactyla*) – Malayan (*M. javanica*) pangolins are shown.

5.2.2 Ancestral population size history

To estimate the past population size of both pangolin species, I extrapolated the observed variations in the diploid pangolins to ancient period using the Pairwise Sequentially Markovian Coalescent (PSMC) model. The principle behind this is the availability of a more diverse mate choice introduces more unique variations and hence the ability to predict population history.

My prediction showed that the *M. javanica* ancestors coalesced at least 5 MYA while *M. pentadactyla* coalesced at least 2 MYA. Interestingly, I observed an inverted population trends between the *M. javanica* and *M. pentadactyla* during the Middle Pleistocene (Figure 5.7). The Marine Isotopic Stage 11 (420-360 KYA) is known as one of the warmest interglacial event in the past 500 KYA, and effective population size, N_e were approaching the maxima and minima for *M. javanica* and *M. pentadactyla*, respectively during this period in Middle Pleistocene (728-126 KYA). The paleoclimatic events during Middle Pleistocene could have provided better genetic fitness for *M. javanica* and thus its rise in their N_e compared to *M. pentadactyla*. Based on these trends, it appears that both species could have different metabolic demands that were affected by environmental stress. It was also interesting to note that the decline in N_e from 100 KYA to 10 KYA coincides with the Late Pleistocene extinction events of other land mammals in other parts of the world.

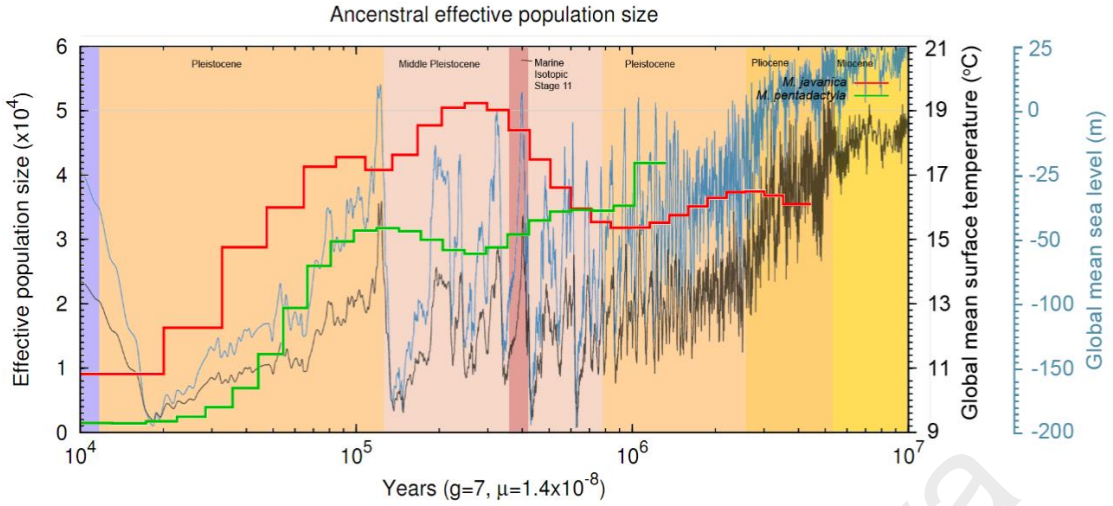

Figure 5.7 Estimated population size history for both *M. javanica* and *M. pentadactyla*.
The X-axis represents time in years in log scale. The first y-axis on the left shows the effective population size scaled to 4μN. On the right, the two colour coded y-axis shows the global mean surface temperature (black) and sea-level data (blue). The global mean sea level is expressed in relation to current sea-level. The paleoclimatic data was obtained from a previous study by Hansen and co-workers (2013).

5.3 Summary

In downstream analysis of pangolin genomes, I detailed out the annotation of repetitive elements including segmental duplication and genes along with a comparison of mean intron length across closely related species. In regards to evolutionary analysis, we estimated the divergence time and applied it to a population genetics model to infer ancient population size of pangolin ancestors. In the next chapter and Section 6.1.2, more information on phylogenetic analysis is given with the assembly of mitochondrial DNA of the pangolins.

CHAPTER 6: ASSEMBLY AND ANALYSIS OF PANGOLIN MITOGENOMES

6.1 Background

While the previous chapters dealt with the analysis of the nuclear genome, here I extend the genome analysis by further analysing the mitochondrial component of the genome or the mitogenome.

6.1.1 General characteristics

The size of the assembled mitogenomes of *M. javanica* MP_PG03-UM and *M. pentadactyla* MPE899 were 16,593 bp and 16,577 bp, respectively (Figure 6.1). Both pangolin species shared 89% genome sequence identity and identities with the mitogenomes of other pangolin species (*M. pentadactyla*†, *M. javanica isolate T298*, *M. tricuspis isolate GLC14*, *M. temminckii isolate T371*, *M. tetradactyla*††, *M. temminckii*) used in this study is also shown in Figure 6.1. In general, the gene number, type and order of the mitogenomes are similar to previously sequenced mitogenomes of other pangolin species. Both mitogenomes of *M. javanica* and *M. pentadactyla* have 13 known protein-coding genes (PCGs), 22 transfer RNA (tRNA) genes, two ribosomal RNA (rRNA) genes, and a non-coding control region (CR). The 13 PCGs in order of appearance are *NAD1*, *NAD2*, *COX1*, *COX2*, *ATP8*, *ATP6*, *COX3*, *NAD3*, *NAD4L*, *NAD4*, *NAD5*, *NAD6*, and *COB*.

The mapped sequencing reads of the assembled mitogenome of *M. javanica* MP_PG03-UM and *M. pentadactyla* MPE899 represented a high mitogenome coverage of approximately 28,000X and 14,000X, respectively. The presence of such large number of sequencing reads requires alternative assembly strategy as typical genome assemblers may fail to derive the optimal assembly. Thus, the assemblies showed that the iterative

mapping and assembly was suitable and successful in assembling the pangolin mitogenomes in this study.

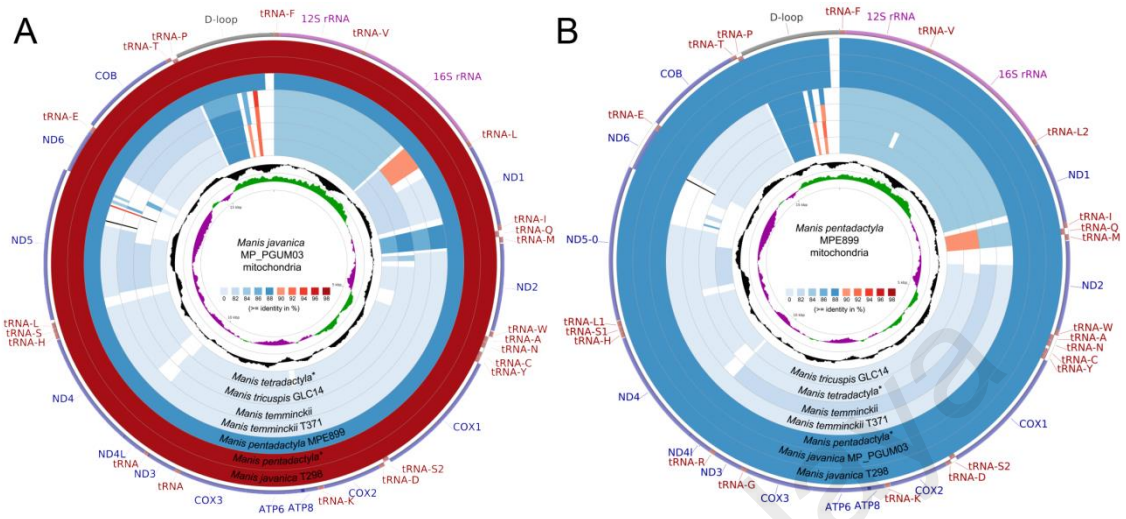

Figure 6.1 Comparison and annotation of *M. javanica* and *M. pentadactyla* mitogenome.
Visualisation of the mitogenome organisation of (A) *M. javanica* MP_PG03-UM, and (B) *M. pentadactyla* MPE899. The outer most rings represent the annotations in the H-strand and subsequently the L-strand. The next seven rings are the alignment of the studied mitogenome with colors shown accordingly with identity as in the legend. The GC content is shown as a black histogram while the GC skews are represented by the green and purple for the heavy and light strands respectively. Arrows represent the orientation of gene transcription and the innermost ring provides the coordinate for the gene location. The pangolin mitogenomes with * are likely misclassified.

6.1.2 Phylogenetic inference

To study the phylogenetic relationship between the pangolin species and to confirm the taxanomic position of the newly assembled sequences, a phylogenetic tree was constructed using available *CYTB* gene sequences from the NCBI GenBank. As anticipated, the assembled mitogenomes of the *M. javanica* MP_PG03-UM and *M. pentadactyla* MPE899 species were placed in its associated species clusters (Figure 6.2). The placement also helps to further confirm the precise identity for the pangolin species that I used in this whole PhD project of the assembled nuclear genome.

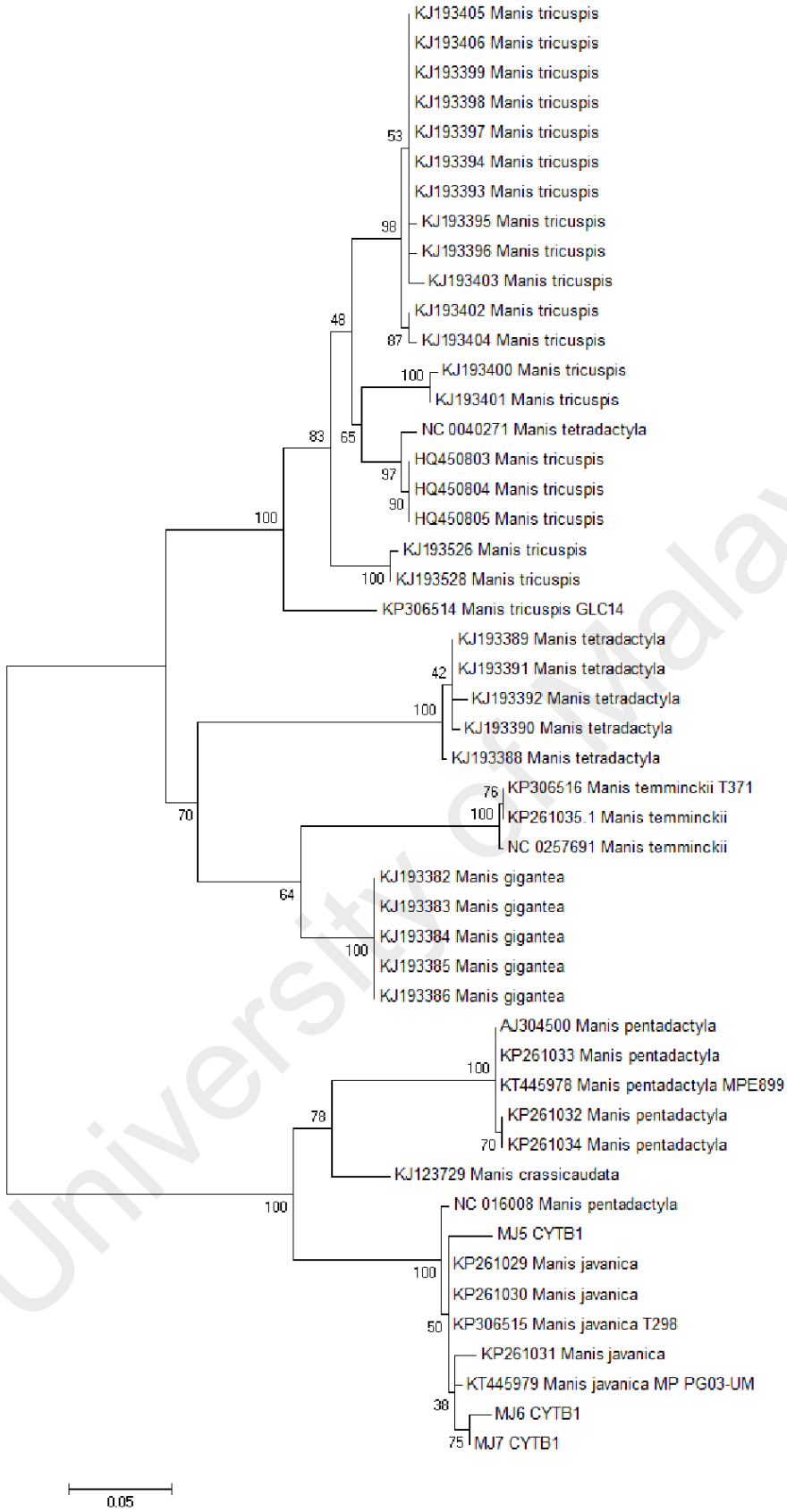

Figure 6.2 Phylogenetic reconstruction using *CYTB* genes.
The phylogenetic tree includes *CYTB* gene sequences (MJ5, MJ6, MJ7) from three wild *M. javanica* in Malaysia which were obtained from the internal sequence database of National Wildlife Department, Malaysia.

Besides that, the findings concurred with previous reports on misidentification of two mitogenomes within the African and Asian pangolin species. Hassanin and co-workers (2015) found two misclassified complete mitogenomes; *M. javanica* and *M. tricuspis* were mistakenly identified as *M. pentadactyla* (NC016008) and *M. tetradactyla* (NC004027), respectively. To further confirm the taxonomic position of the pangolin mitogenomes, a robust and reliable phylogenetic tree was constructed using whole mitogenome sequences of from different pangolin species. The whole mitogenome-generated tree further supported the view that the *M. tetradactyla* (NC004027) and *M. pentadactyla* (NC_016008) were wrongly classified as we observed in the above *CYTB*-derived trees (Figure 6.3).

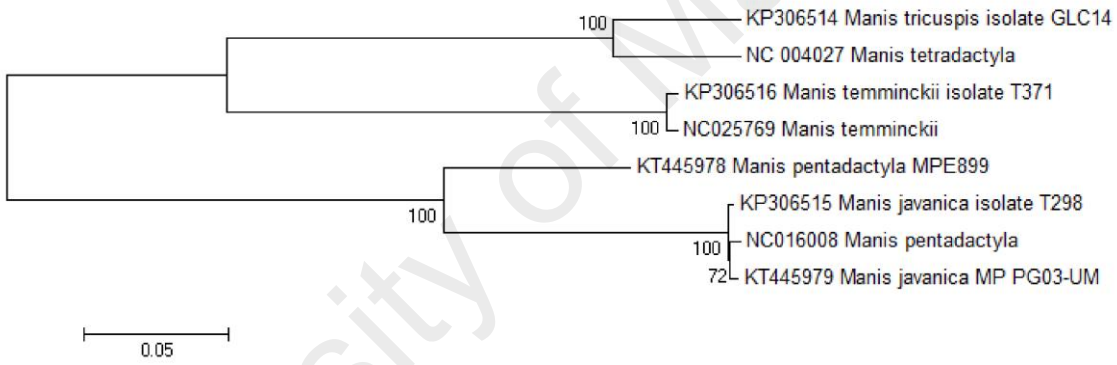

Figure 6.3 Phylogenetic tree construction using whole mitogenome of pangolins
The complete mitogenomes of available pangolin species were aligned with MAFFT software and the phylogeny was reconstructed using Maximum Composite Likelihood (MCL) distance with 1000 bootstrapping replications.

Therefore, the *M. pentadactyla* mitogenome sequence that I generated in this study is important and can be the first new reference mitogoneome for *M. pentadactyla*, replacing the misclassified *M. pentadactyla* mitogenome sequence (accession NC016008) in the Genbank database.

6.1.3 Transfer RNA genes

All 22 tRNAs complement typical of metazoan mitochondria were found to be present (Figure 6.4). The tRNA has as an average size of 68bp and a sum of 1496bp (9.01%) and 1497bp (9.03%) for *M. javanica* MP_PG03-UM and *M. pentadactyla* MPE899, respectively. Most tRNAs exhibit the typical clover-leaf secondary structure except *tRNAS1* which exhibited a loss in dihydrouridine (DHU) arm. The loss of the DHU arm is commonly observed in other metazoan mitogenomes (Wolstenholme, 1992).

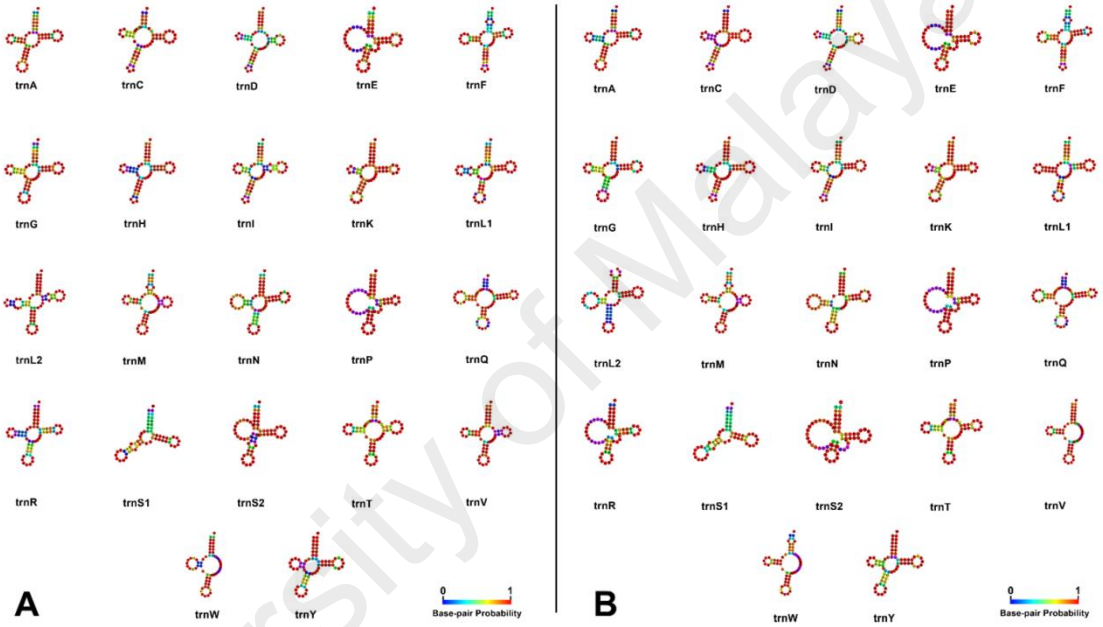

Figure 6.4 Predicted secondary tRNA structures of *M. javanica* and *M. pentadactyla* mitogenome.
(A) *M. javanica* (B) *M. pentadactyla*. The base-pairs are coloured based on base-pair probabilities while for unpaired regions the colour denoted the probability of being unpaired. The structural prediction diagrams was generated using RNAfold (Hofacker, 2009).

6.1.4 Codon usage

In this analysis, I found that the patterns of gene level codon usage were different between the Asian and African pangolin species, based on the clustering results observed in the PCGs heat map (Figure 6.5). The Asian and African pangolins formed major distinct branches when clustered. It was also interesting to note that clustering analysis based on PCGs RSCU values grouped the species in a similar way as observed in the phylogenetic

tree. It suggested that the PCG codon usage patterns seemed to be similar in closely related species, as previously reported (Sharp et al., 1988, p. 1988). However, the cluster of these two groups was not maintained when compared separately with each 13 gene-wise plots (Appendix 2). The differences observed in the gene-wise cluster could be due to the respective codon usage bias of genes reflecting a mutation-selection balance at a point determined by the strength of translational selection on that gene (Sharp, Tuohy, & Mosurski, 1986).

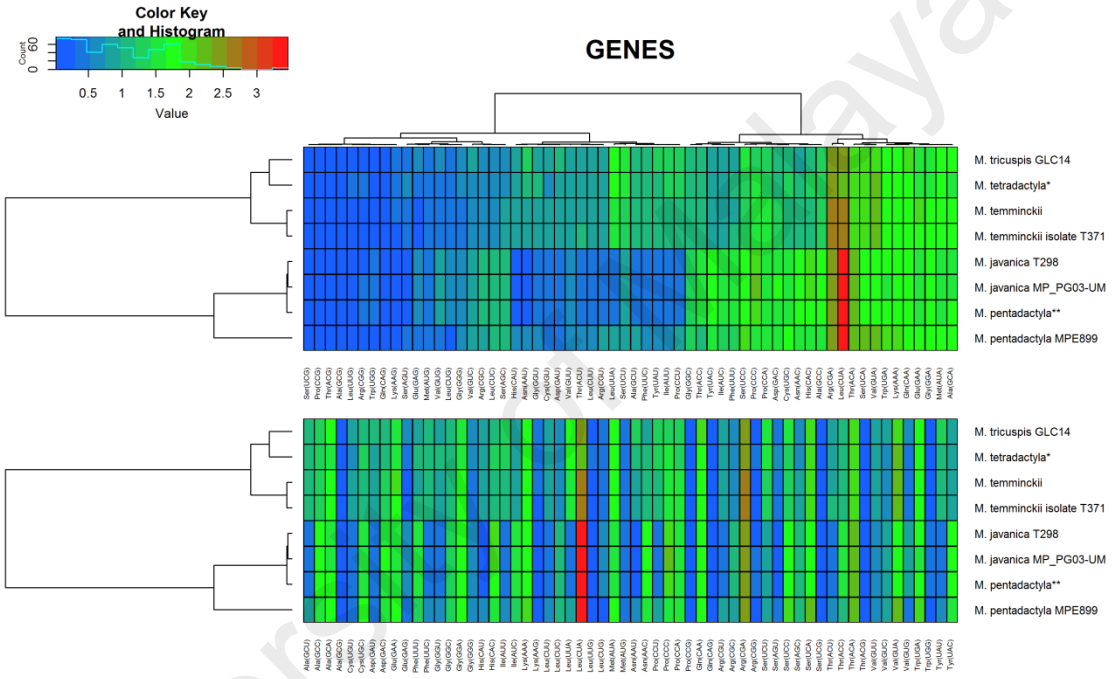

Figure 6.5 The codon usage of all concatenated protein coding genes is shown as a clustered heatmap.
(A) Clustering was performed based on Minkowski distance both on pangolin species and codon. (B) The columns were ordered based on the amino acids which codons code. Figure 6.5-B enables one to quickly identify the common codons which are used by the organism compared to other degenerate codons. *M. pentadactyla** (inferred as *M. javanica*) and *M. tetradactyla*** (inferred as *M. tricuspis*) are misclassified sequences.

6.2 Summary

In this chapter, I have presented the first *M. pentadactyla* mitogenome and sequenced and compared with the mitogenome of its closest relative, *M. javanica*. The addition of *M. pentadactyla* and *M. javanica* mitogenome sequences can serve as reference mitogenomes or as useful resources for future pangolin studies as well as for use in legal and forensic framework for prosecuting illegal pangolin traders.

CHAPTER 7: IDENTIFICATION AND ANALYSIS OF LNCRNA

7.1 Overview

While pangolins exhibit a wide range of adaptations to their unique characteristic, it is useful to compare the identified lncRNA in pangolins to expand and validate known lncRNA annotations in humans and other organisms. Generally, the identification and study of lncRNAs is relevant to not only pangolin but also human biology and disease since they include an extensive set of mostly unexplored functional component of the genome (Mattick 2009; Ponting et al. 2009). By and large, some lncRNAs have been shown to affect human diseases, but these studies are limited by the lack of lncRNA annotations. Therefore, the identification of high-quality catalogues of lncRNAs and its expression in tissues is an important prerequisite to better understand the function of therapeutic and regulatory actions of this class of ncRNAs.

7.2 LncRNA identification

In order to systematically identify or catalogue, the expressed lncRNA repertoire of the pangolin, I utilized eight tissue specific cDNA libraries of *M. javanica* from our RNA-seq study (Aini et al., 2016). These libraries were constructed in a strand-specific manner to yield a sequenced read length of 100bp using the Illumina HiSeq 2000 platform. The sequencing produced more than 373 million paired-end reads in total and each sequenced tissue libraries were independently assembled using a reference annotation based transcript (RABT) assembly method with the reference gene annotation obtained through this study in section 5.1.2. Finally, each resulting assembly were merged using Cuffmerge to generate a consensus transcriptome (n=117,058) across the eight tissue samples. In our previous pangolin transcriptome sequencing study (Aini et al., 2016), a *de novo* transcript assembly strategy without a reference yielded 106,709 transcripts prior to clustering. The

RABT method yielded 10,349 transcripts more than the *de novo* assembly method. Thus, the new RABT-based assembled transcriptome was used a starting point in the identification of pangolin lncRNA.

An in-house designed integrated pipeline was developed to stringently filter down the total 117,058 pangolin transcripts to a final high confident set of lncRNA (n=7,422). Similar methods have been successfully used in previous studies in identifying putative stringent lncRNA (Gaiti et al., 2015; Tilgner et al., 2012). A broad range of class codes 'i', 'o', 'u', 'j' and 'x' in the initial step was chosen to be retained representing transcripts that are intronic, overlapping exons, unknown, have a splice junction overlap, and exonic overlap in the opposite strand respectively. Subsequently, transcripts were filtered based on three main factors, i.e., length of more than 200 nucleotides, multi-exonic characteristic, and non-intersection with known protein coding exon. Finally, the probability of the transcript being a protein was addressed using state-of-the-art coding potential predictors and homology search against the Pfam database. The reduction in transcript numbers in each step of the pipeline is shown in Table 7.1. The resulting set of transcripts is considered to be high confident lncRNAs. I have identified a total of 7,422 lncRNA transcripts at the end of running the whole analysis pipeline.

Table 7.1 Transcript counts at every stage of the systematic computational pipeline for the discovery of pangolin lncRNA transcripts

No	Pipeline Filters	Transcript count
0	Initial number of transcripts	117058
1	Retain cuffcompare classes – i, o, u, j, & x	108746
2	Retain transcripts >200nt in length	71826
3	Filter single exons transcripts	46211
4	Filter protein coding potential	10042
5	Filter Pfam hits	8842
6	Filter protein-coding sense exon intersect	7422

7.2.1 Comparison with protein coding genes

Generally, there are specific signatures of lncRNA previously established in other studies. I found that the pangolin lncRNA had lower number of exons and on average possess 2.8 exons per transcript compared to 7.2 exons in protein coding genes (Figure 7.1). However, the mean exon length of the pangolin lncRNA were found to be higher (539bp) compared to its protein-coding genes (261bp) (Figure 7.2). The gene lengths are also highly dissimilar with lncRNA (mean=15kbp) generally shorter on average to protein-coding genes (mean=21 kbp) as shown in Figure 7.3. In summary, the lncRNA transcripts are generally shorter than the protein-coding genes and possess lesser exons, however their exon lengths appear longer than coding genes, probably due to high number of shorter exons present in the protein-coding genes in pangolins.

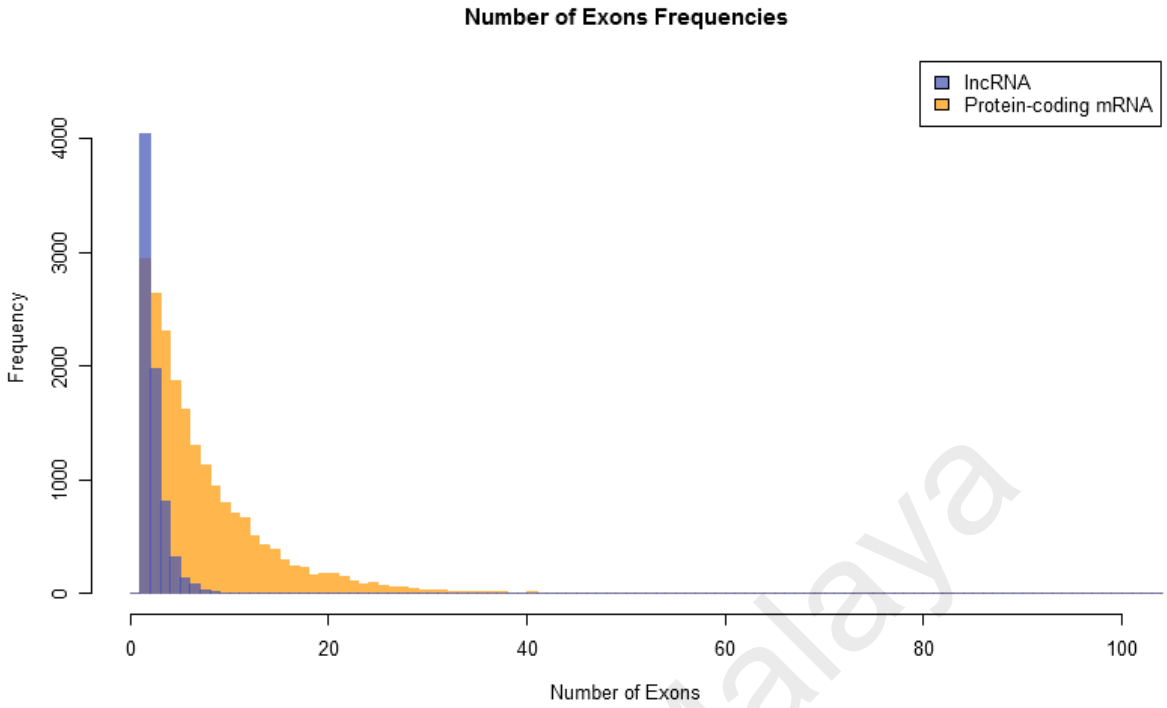

Figure 7.1 Exon number frequency
Frequency chart showed the number of exons contained in each lncRNA and protein-coding mRNA

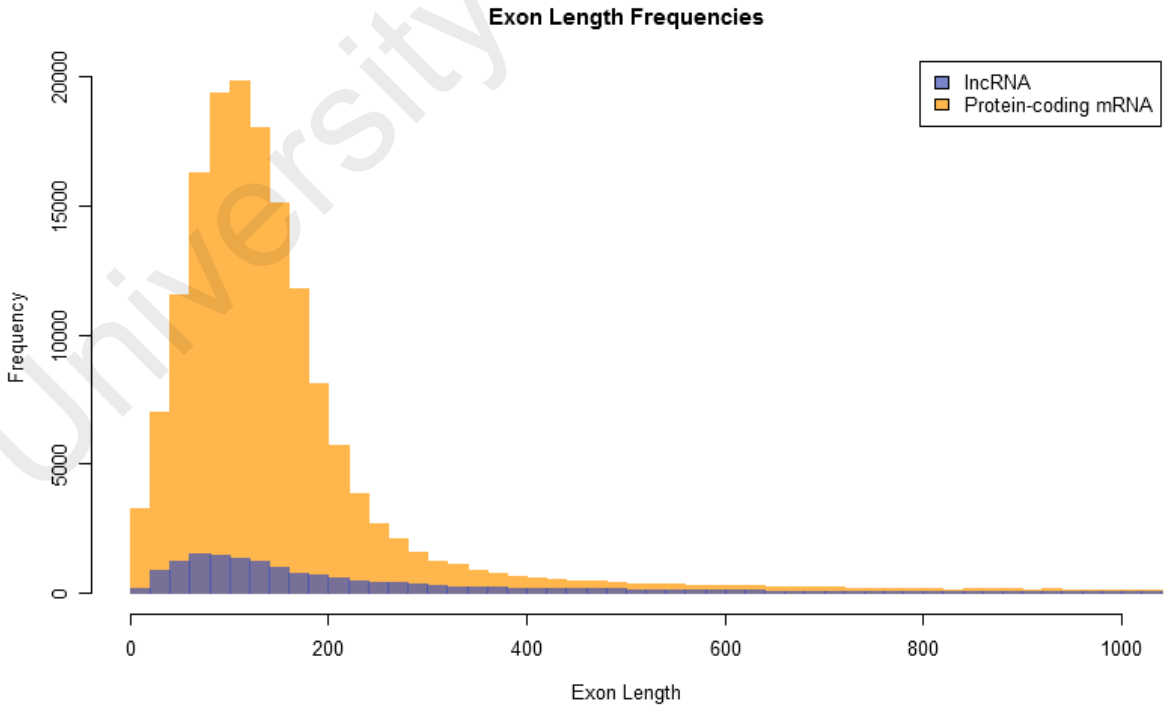

Figure 7.2 Exon length distribution frequency
The frequency of lncRNA and protein-coding mRNA containing a particular exon-length is shown here.

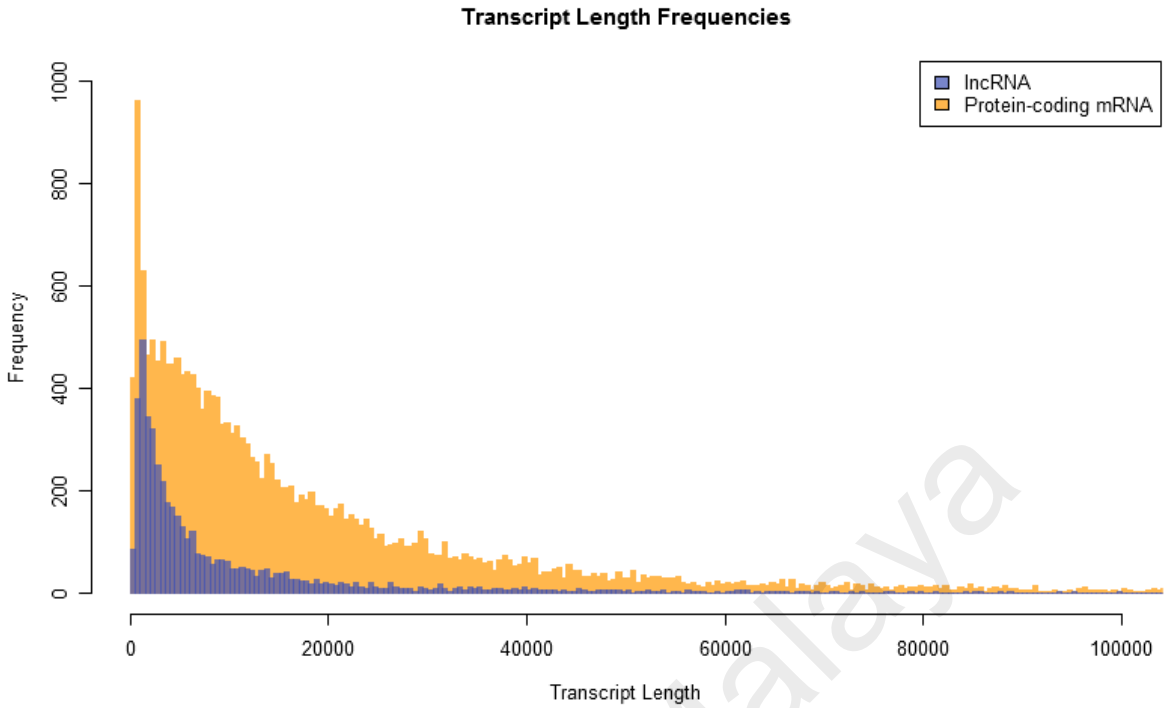

Figure 7.3 Transcript length distribution frequency
Histogram showed the length of transcript and its frequency of both lncRNA and protein-coding mRNA

7.2.2 Classification of lncRNA genes

The list of predicted lncRNA genes were further classified based on their localisation in relation to the protein-coding genes in the pangolin genome (Figure 7.4). Primarily, lncRNAs were classified into intergenic and genic subclasses. Wherever protein coding genes annotation was absent in the genome sequence, intergenic lncRNAs present were labelled as orphan. Otherwise, the classification of the lncRNA were adapted from schemes used in previous studies (Derrien et al., 2012). For instance, the lncRNA that do not overlap any genic regions were classified as intergenic lncRNA also known as lincRNA which represent the largest class in this study's catalogue (6123). The remaining number of transcripts were that which intersected genic regions and accounted for exonic and intronic intersects of 589 and 710 respectively (Table 7.2).

Table 7.2 Sub-classification of high confidence pangolin LncRNAs.

Pangolin LncRNA (7,422)							
Intergenic (6123)				Genic (1300)			
Orphan	Convergent	Divergent	Same Strand	Exonic (589)		Intronic (710)	
2874	683	1072	1494	S	AS	S	AS
				-	589	441	269

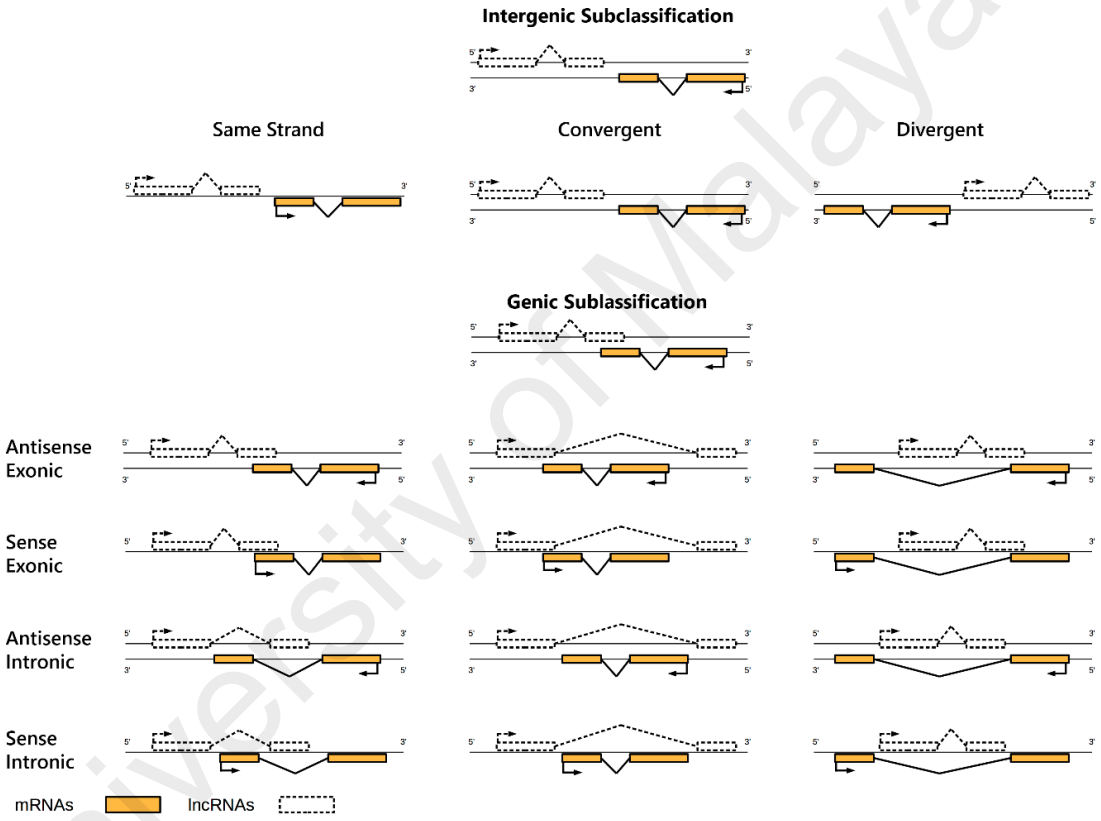

Figure 7.4 Classification of lncRNA based on localisation in relation to coding genes

7.2.3 Repeat Content in lncRNAs

Repeat annotations on pangolin lncRNA transcripts at the level of classification were compared by percentage content (Figure 7.5). Besides that, the repeat content of the protein-coding mRNA and lncRNA transcripts were compared (Figure 7.6). As expected, intergenic lncRNA harbours the most number of repeats compared to the antisense (AS) exonic and intronic lncRNAs (Figure 7.5). However, the content of repeat classes was

proportionally similar between each lncRNA classes. While approximately 50% of the lncRNAs contain repeat sequences, typical for a coding gene less than 5% repeat content is observed. It is expected that coding genes have low repeat percentage and contain minimal repetitive regions to optimally produce functional proteins. However, in this catalogue of lncRNA they display high repeat content as found in previous similar studies for other organisms (Gaiti et al., 2015; Kannan et al., 2015; Kapusta et al., 2013; H.-Q. Liu, Li, Irwin, Zhang, & Wu, 2014).

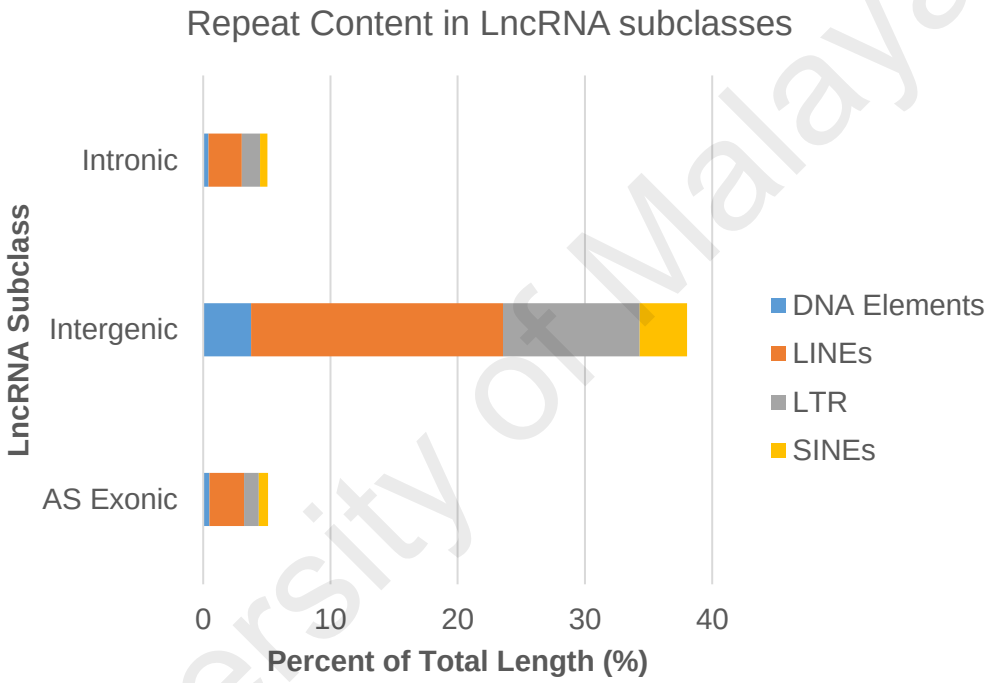

Figure 7.5 Repeat proportions per lncRNA classification
Each type of lncRNA class are shown in terms of repeat content of common classes of DNA elements, LINEs, LTR and SINEs.

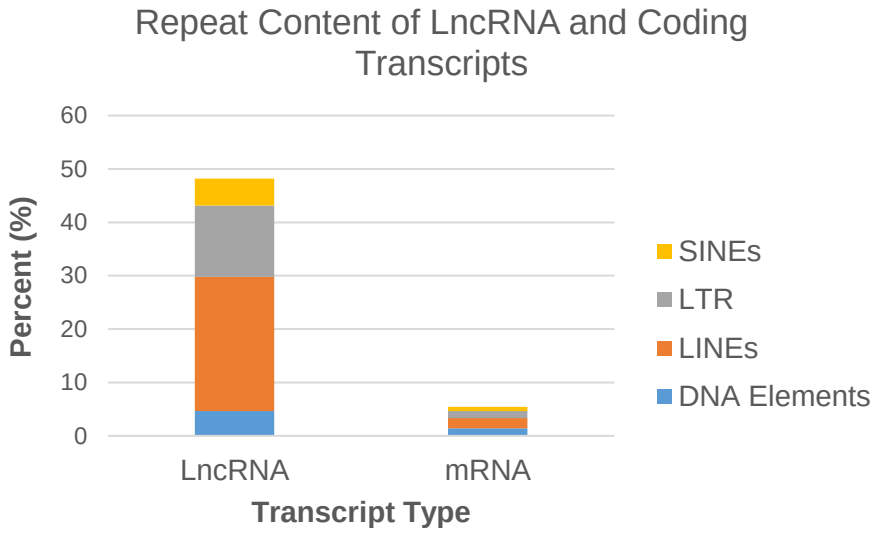

Figure 7.6 Proportion of repeats in lncRNA and protein-coding mRNAs
Comparison of repeats and their classes in percentage between lncRNA and protein-coding mRNA transcipts

7.2.4 Expression patterns of pangolin LncRNA

Previous studies have concluded that lncRNA could play important roles in development biology of organisms (Fatica & Bozzoni, 2014). Moreover, several studies have noted that the majority of the expression patterns of lncRNA are likely to follow a tissue specific trend (Derrien et al., 2012). Using the Jensen-Shannon divergence to cluster the expression values, distinct clusters of lncRNA expression in pangolins were visualized in the heatmap (Figure 7.7). This clusters mean there could be a pattern of tissue specificity in pangolin lncRNA as well. However, the number of lncRNA participating in tissue specific expression (where J-S threshold > 0.5) is relatively low (37.1%) compared to other studies which exhibit more than 70% of the catalogued lncRNA. Non-lncrna transcripts that show tissue specificity is at 33.7% and only slightly lower than lncRNA transcript.

Figure 7.7 A heatmap of showing the expression patterns of lncRNA genes across different pangolin tissues.

Each brown line describes a high level of expression. Clustering has been applied to group lncRNA that are highly expressing by tissue type. Several distinct clusters can be observed that show distinct expression pattern by tissue.

7.3 Summary

In this chapter, I have annotated a high-quality set of non-coding RNAs in pangolin which are difficult to identify with the genome or transcriptome alone. By integrating the set of two genome and transcriptome data that we had, I was able to systematically mine for lncRNA. A total of 7,422 lncRNAs were found that meets the stringent factors that are used to detect lncRNAs.

CHAPTER 8: DISCUSSION

8.1 Summary of the study

In this study, I have successfully sequenced and assembled the first nuclear genome sequence of *M. javanica* genome and compared with the genome of its closely-related species, *M. pentadactyla*. Segmental duplication counts and its coverage were tabulated and the repeat landscape over a phylogenetic distance was also analysed. Besides that, repeat composition that was examined in relation to genic, intronic and intergenic regions uncovered the absence of tRNA-SINE activity. When I annotated the genes in the genomes, both pangolin genomes presented the lowest mean intron length among the mammals screened possibly due to the depletion of the SINE family. In ancestral population size analysis, population size followed an inverted trend in relation to *M. javanica* ever since ancestors of *M. pentadactyla* coalesced 2 MYA until 100 KYA.

As per read characterisation performed early in the study, an indication of foreign DNA was found. Upon further investigation, several cases of colonisation by *Burkholderia fungorum* in the pangolins were found. Later, it was found that the *M. javanica* brain and lung were also colonised by this pathogen.

8.2 Genome size differences in *M. javanica* and *M. pentadactyla*

While the genome size prediction showed that the *M. pentadactyla* (2.69 Gbp) genome size is larger than *M. javanica* (2.45 Gbp), the final assembly was 2.2 Gbp and 2.55 Gbp respectively raising question to the discrepancy in *M. pentadactyla* genome size. While it should be noted that the *M. pentadactyla* was sequenced at low coverage, therefore, its genome assembly was fragmented and may not represent the complete genome. Despite this limitation, segmental duplications were identified in a higher proportion in *M.*

pentadactyla (44.8 Mbp – 1.66%) compared to *M. javanica* (36.2 Mbp – 1.48%). This in part could explain the predicted larger genome size in *M. pentadactyla*. Sometimes a large portion (up to 30%) of some genomes could be represented by repeat or transposable element sequences and recent repeat activity could be seen to mark the difference in diverged species. In my repeat landscape analysis, it was observed that there was no particular recent repeat activity that was enhanced particularly in *M. pentadactyla* compared to *M. javanica*. Besides that, previous karyotyping analysis have shown that *M. pentadactyla* has a higher diploid number (40) compared to *M. javanica* (38) (Nie et al., 2009) that could have arisen from chromosome fusion thus inflating the genome size in *M. pentadactyla*. Moreover, there are variation in diploid number in *M. pentadactyla* that have been previously reported ranging from 36 to 42. Taken together, genome level rearrangements and duplications could be major contributors to the difference in genome size between the two pangolins instead of genome expansion through repeat activity.

8.3 Lack of tRNA-SINEs and lower mean intron length

Unexpectedly, I found that pangolins have significantly low percentage of SINE repeats compared to their closely related Carnivoran genome such as panda, dog, tiger and cat. My data clearly indicates that the low percentage of SINE repeats in pangolins is likely due to the lack of tRNA-derived SINEs in pangolins while harbouring ancient copies of SINE MIR family of repeats. A possible explanation for the lack of tRNA-derived SINEs in pangolins is perhaps the presence of a molecular system which excises and prevents tRNA-derived SINEs from propagating within their genomes as similarly suggested for the chicken genome (Frésard et al., 2014; Hillier et al., 2004). As previously reported, the host genomes are known to evolve complex mechanisms to defend its genome from spurious propagation of transposable elements (Cantrell, Scott, Brown, Martinez, & Wichman, 2008; Erickson, Cantrell, Scott, & Wichman, 2011; Platt II & Ray, 2012).

Normally, SINE elements are observed to show bursts of activity followed by periods of quiescence for unknown reason (Hormozdiari et al., 2013). The lack or absence of tRNA-SINEs has not been observed before in other mammalian genomes when MIR SINE family is present. Nevertheless, the absence of tRNA-SINE could have a much more important role that is yet to be known in pangolins.

I believe that the excision or absence of SINE elements might have caused the average intron length of the pangolin genome to be generally smaller in comparison to other closely-related mammals. Nevertheless, such profound changes at the genomic level is thought to have important consequences to the genetic fitness of the species.

For instance, two independent comparisons between flying and nonflying sister groups of avians showed the presence of smaller intron sizes in the volant species suggesting that metabolic rates associated to power flight could be associated to a reduction in genome or intron size (Hughes & Hughes, 1995). Moreover, a rigorous study that addresses previous shortcomings further confirmed the association of flight and genome size to intron lengths (Qu Zhang & Edwards, 2012). For pangolins, the metabolic demand of having limited prey availability due to its myremecophagous nature could be hypothesized for its shorter intron size. The typical behaviour of pangolins curling into a ball when threatened could also explain pangolins naturally attempting to conserve energy when dealing with a fight or flight scenario. Therefore, in a general sense, I postulated that the absence of tRNA-SINE family and a lower average intron size could be related to the energy conservation in pangolins.

8.4 Presence of *B. fungorum* in *M. javanica*

Unexpectedly, our contamination screening of pangolin assembly revealed the presence of *B. fungorum* sequences, suggesting the colonisation of this bacterium in the cerebrum and cerebellum (but not in liver) of the female *M. javanica* used in this study. PCR assays confirmed the presence of this bacterium in the two sterile tissues and also in the lung and blood tissues. Interestingly, PCR analyses using blood samples of seven individual *M. javanica* also showed the presence of *B. fungorum* present in more than half of the pangolins that were examined, supporting the view that is unlikely an isolated case or the contamination in the organ samples during wet laboratory experiments. Perhaps pangolins could be naturally prone to *B. fungorum* infection due to their frequent exposure to soil where *Burkholderia* species are likely abundant (Salles, van Veen, & van Elsas, 2004). With the recent finding that an immune-related gene (*IFNE*) that is important for skin and mucosal immunities is pseudogenised in pangolins (this gene is intact in other 71 mammals examined in this study), suggesting that pangolins might have poor skin or mucosal immunity (Choo et al., 2016). We cannot rule out the possibility that the *B. fungorum* might manage to evade host-defense mechanism and colonise pangolins easily especially when pangolins are under stress. It has been reported that *B. fungorum* present in cerebro-spinal fluid of a woman (Coenye et al., 2001). Moreover, other report has shown that *Burkholderia* species being capable of direct infection of the brain via olfactory receptor neurons extending from the nose (John et al., 2014). Therefore, the route of entry of *Burkholderia* could be mainly the nasal cavity or lungs. This is reflected by the PCR assays where mainly brain and lung tissues showed the presence of *B. fungorum* DNA in *M. javanica* used in this study. The possibly poor immunity of pangolin (eg. due to the loss of *IFNE* gene) and the presence of pathogens must be taken into account during captive breeding efforts where stress can overwhelm the immune protection against such colonisers. On another note, it is possible that the brain and other

tissues may express important genes that protects pangolins from adverse effects of bacterial infection. Generating gene expression data could be an important resource where new therapeutic transcripts can be found.

8.5 The lncRNome of pangolins

As part of this study, I have identified a comprehensive set of pangolin lncRNA using stranded RNA-seq across eight different tissue samples developed using a systematic filtering pipeline. This is the first catalogue of pangolin lncRNA transcripts in multiple tissue types which resulted in the identification of 7,422 lncRNA. This catalogue represents 7,422 multi-exonic transcripts which are subclassifed into lincRNAs, intronic lncRNA and antisense exonic lncRNA. These pangolin lncRNA share many of the characteristics of lncRNA genes from other species such as being relatively short and lower in exon numbers. The expression levels of lncRNA are also generally lower than protein coding transcripts. Besides that, the expression is pervasive throughout the genome and indicate a major presence in the intergenic regions.

A recent study suggested that transposable elements help shape the origin and diversity of the lncRNA repertoire. Transposable elements form majority of the content found in lncRNA genes as per repeat analysis. On average, the pangolin lncRNA is covered up to 50% of its length by repetitive elements in this study. This is comparable to previous results, yet specifically the pangolin lincRNA possess the highest repeats compared to human, mouse and zebrafish lincRNA sets (Kapusta et al., 2013).

Many lncRNA participate in regulating diseases, and vascular functions (Boon, Jaé, Holdt, & Dimmeler, 2016). It can be a resource for therapeutic potential in regulating disease. For instance Nuclear enriched abundant transcript 1 (NEAT1) is a lncRNA

widely expressed in a wide range of tissues (Clemson et al., 2009; Sasaki, Ideue, Sano, Mituyama, & Hirose, 2009), highly expressed in the brain when facing certain neurodegenerative conditions or viral infections (R. Johnson, 2012; Quan Zhang, Chen, Yedavalli, & Jeang, 2013). Moreover, in knock-down models, the progression of disease is much faster. Therefore, the lncRNA repertoire may be an untapped target for therapeutics.

8.6 Dwindling trend in ancient population size history

Using a PSMC model, the ancient population size structure of *M. javanica* and *M. pentadactyla* ancestors were predicted. The prediction showed inverted population trends of *M. javanica* and *M. pentadactyla* during the Middle Pleistocene (Figure 5.7). Particularly during Middle Pleistocene (728-126 KYA), after Marine Isotopic Stage 11 (420-360 KYA) which is the warmest interglacial period 500 KYA, the effective population size, N_e were approaching the maxima for *M. javanica* and in the contrary the minima for *M. pentadactyla*. I postulated that the geological and paleoclimatic conditions during Middle Pleistocene might have provided better genetic fitness for *M. javanica* and thus its rise in their N_e compared to *M. pentadactyla*. Moreover, it is possible that both species could have different metabolic demands that were affected by environmental stress. Recently, an associated study also showed that pangolins have poor immunity (Choo et al., 2016). This could predispose the pangolin to diseases that are disseminated during particular climate changes. This information about environmental stressors is useful in conservation efforts of pangolins. By anticipating pangolins' sensitivities to environment, climate regulation can be part of the artificial habitat that conservationist need to adopt to have a successful captive breeding and protection program.

8.7 Importance of using mitogenome in identifying the species of study

Apart from the nuclear genomes of pangolins, I have also characterized the NGS data-derived mitogenomes of *M. javanica* and *M. pentadactyla*. Mitochondrial markers are important resources for identification of pangolin species. Recent reports of misclassification was confirmed in the course of phylogenetic analysis in this project and it can be a concern (Hassanin et al., 2015). Due to close morphological resemblance among the Asiatic pangolin species, animal samples could be easily misidentified. Thus, it is imperative that confirmation of mitochondrial barcode sequences such as *CYTB* are checked for similarity against a reliable database. In this study, two whole mitogenomes were provided as an important genetic resource for researchers and law enforcement officers dealing with animal forensics of illegal trade in future. For researchers, it is imperative that the species is confirmed before as subsequent analysis may be performed more reliably.

8.8 Challenges and improvement of pangolin genome assemblies

Large genome sequencing projects could have a computational cost depending on the size of the genome being sequenced. For instance, any genomes larger than 2 Gbp usually have computational memory bottlenecks when assembling large NGS raw data (Zerbino & Birney, 2008). Before performing an assembly, it is useful to be able to understand the characteristic of the genomes. In case it is a novel species where reference sequences are unavailable, a genome project's first bioinformatics step, the genome assembly, often starts mostly on trial and error in order to optimize the final genome assembly.

As outlined in Table 2.4, the choice of tools used for assembly will depend on the sequencing platforms and the read length that were used besides the computational resources available (Ekblom & Wolf, 2014). There are many genome assemblers that

have been developed, for example, SOAPdenovo2 (R. Luo et al., 2012, p. 2), ALLPATHS-LG (Gnerre et al., 2011), Velvet (Zerbino & Birney, 2008), and SGA (Simpson & Durbin, 2010). Although I tried to assemble with Velvet and ALLPATHS-LG, but failed due to memory usage exceeding the available limit of available random access memory (RAM) of our in-house cluster computer. Therefore, I used SOAPdenovo2, CLC Assembly Cell (CLC bio, Aarhus, Denmark) and SGA which need lower requirement of memory. SOAPdenovo2 assembler can allow the graph construction step to run under a sparse mode in order to reduce the computational memory demand. Besides that, the proprietary CLC Assembly Cell program and SGA are able to function within the limit of the commodity cluster used in this study. Thus, the three assemblers could be assemblers of choice for small laboratories using commodity cluster when dealing with genome projects.

While the basic goal of any genome project is to get the longest and most contiguous DNA sequence possible reflecting chromosomal structure, however, achieving it is challenging in many different levels. Although Illumina-based NGS platforms are commonly used in large-scale genome projects, but this technology has limitations. For instance, this technology only generates short reads which may be not good for assembling repetitive regions in the genomes. Furthermore, genomes that are highly heterozygous (> 0.5%) may also pose a problem by introducing ambiguity in the assembly phase due to short reads.

To enhance the current draft pangolin assemblies, it is possible to try new genome assemblers that are suited to the heterozygous genomes such as redundans (Pryszcz & Gabaldón, 2016) and Platanus (Kajitani et al., 2014). Besides that, we could also try to use additional large mate-air libraries e.g. 10kb, 15kb, 20kb and so on, which can help to

scaffold the fragmented contigs or filling the gaps. sequencing is a good solution that can be explored. Moreover, a long-read sequencing technology such as Pacbio can also be used to further enhance the current pangolin assemblies. This technology can generate long reads that are useful to fill the gaps, elucidate repeat rich regions, and for scaffolding.

8.9 Impact of the study

Prior to undertaking this study, there was little direction to work on the genetics of pangolins. One of the main contribution of the study, is to provide the first reference genome or genomic resources to the pangolin research community. Secondly, in this study I provided salient information that characterizes the genome. More generally, findings of genome size estimations, annotations of transposable elements, gene and lncRNA will be a basis for quicker future comparative analysis. Specifically, the annotated set of genes for both *M. javanica* and *M. pentadactyla* can be used as a starting point for various genetic and medicinal research in future. Besides that, the annotated genes found in pangolin could also assist in improving the annotation by providing support to the existing annotations in the human genome project. In evolutionary analysis, the mutation rate unique to the pangolin species using the whole genome data were calculated. This value is useful to guide future research in areas such as population genomics and was used for latest pangolin research in Singapore (H. C. Nash, personal communication, July 22, 2016). In this study, it was used in properly delineating the timeline of the historical population size estimate plot. In terms of systematics, the associated mitogenomes was provided and served as a DNA barcode to identify local species and can be applied in legal and forensics to assist in prosecuting illegal pangolin trade.

8.10 Future Work

8.10.1 Understanding important traits of the pangolin

With the availability of the pangolin genome sequences and annotations, several candidate genes of interest could be subject to detailed examination in future. For instance, these selected genes could perhaps focus on tooth development, keratinous scales and vision. By looking at these traits, one could identify the evolutionary adaptations specific to the genome. Moreover, one may also be able to compare the orthologous genes of different mammals to observe the evolutionary changes. Apart from this, the functional as well detailed pathways that these genes participate in can also be characterized. This will enable us to perform knock-out studies to identify the function of the genes in murine or other animal models.

CHAPTER 9: CONCLUSION

Despite growing interests on the endangered pangolins, research efforts may be limited by difficulty in getting samples and perhaps the lack of research coordination. In such challenging situations, the availability of a genetic resource particularly the whole-genome sequence can serve as an indispensable resource in furthering pangolin research. Here, I have successfully sequenced, assembled and analysed the nuclear genomes and mitogenomes of pangolins, which could contribute positively as reference genomes for pangolin research in future. The pangolin genes that we identified and annotated in this study also provide useful resource for functional studies in future.

Strikingly, I have also reported that pangolins have a noncanonical repeat layout. The transposable elements in pangolin genomes are generally lower than in other Carnivora, suggesting divergence in genome architecture between pangolins and other placental mammals. They also have relatively low proportions of SINE repeats and intronic repeats, mainly due to a relatively smaller number of tRNA SINEs. It is possible that low percentage of intronic repeats have contributed to lower mean intron length in pangolins as I observed in this study.

Moreover, I have also suggested that segmental duplications and other chromosomal rearrangements might contribute to the significant difference between the genomes of *M. javanica* and *M. pentadactyla.* In evolutionary analysis, besides identifying divergence time, I showed that the pangolin population is dwindling and their numbers have been decreasing since late Pleistocene, 100 KYA. Paleo-geoclimatic events could have affected the genetic fitness of the pangolin population, concurring with previous reports

of the fragile nature of pangolins susceptible to respiratory problems when temperatures drop below 21^{o}C (V. T. Nguyen et al., 2010).

Furthermore, I have generated the associated organellar mitogenome sequences of both pangolin species used in this study, serving as an important reference for other studies especially with *M. pentadactyla* mitogenome being a novel sequence in public repository.

By integrating data from RNA-seq, I identified a confident set of expressed lncRNA in *M. javanica* and grouped them into seven spatial categories. Intriguingly, while lncRNA expression patterns are normally tissue-specific, however, it was not apparent in this dataset. The lncRNA identified in this study will serve as a fundamental resource in comparative studies in this newly active field of research.

In conclusion, this study provides new insights into the biology, evolution, and diversification of pangolins. The annotated reference genomes will be invaluable for future studies addressing issues of species conservation, gene function and forensics.

REFERENCES

Aini, M. Y., Tze King, T., Ranjeev, H., Klaus-Peter, K., Wei Yee, W., Agostinho, A., … Siew Woh, C. (2016). De novo sequencing, assembly and analysis of eight different transcriptomes from the Malayan pangolin. *Scientific Reports*.

Alam, M. T., Petit III, R. A., Read, T. D., & Dove, A. D. M. (2014). The complete mitochondrial genome sequence of the world's largest fish, the whale shark (Rhincodon typus), and its comparison with those of related shark species. *Gene*, *539*(1), 44–49. https://doi.org/10.1016/j.gene.2014.01.064

Almendro, V., Carbó, N., Busquets, S., Figueras, M., Tessitore, L., López-Soriano, F. J., & Argilés, J. M. (2003). Sepsis induces DNA fragmentation in rat skeletal muscle. *European Cytokine Network*, *14*(4), 256–259.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Andrews, S., Krueger, F., Seconds-Pichon, A., Biggins, F., & Wingett, S. (2014). *FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics*.

Angeloni, F., Wagemaker, N., Vergeer, P., & Ouborg, J. (2012). Genomic toolboxes for conservation biologists. *Evolutionary Applications*, *5*(2), 130–143. https://doi.org/10.1111/j.1752-4571.2011.00217.x

Arnason, U., Adegoke, J. A., Bodin, K., Born, E. W., Esa, Y. B., Gullberg, A., … Janke, A. (2002). Mammalian mitogenomic relationships and the root of the eutherian tree. *Proceedings of the National Academy of Sciences*, *99*(12), 8151–8156.

Austin, C. M., Tan, M. H., Croft, L. J., & Gan, H. M. (2016). The complete mitogenome of the crayfish Cherax glaber (Crustacea: Decapoda: Parastacidae). *Mitochondrial DNA Part A*, *27*(1), 220–221.

Bagatharia, S. B., Joshi, M. N., Pandya, R. V., Pandit, A. S., Patel, R. P., Desai, S. M., … Saxena, A. K. (2013). Complete mitogenome of asiatic lion resolves phylogenetic status within Panthera. *BMC Genomics*, *14*(1), 572. https://doi.org/10.1186/1471-2164-14-572

Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M., & Eichler, E. E. (2004). Analysis of Segmental Duplications and Genome Assembly in the Mouse. *Genome Research*, *14*(5), 789–801. https://doi.org/10.1101/gr.2238404

Benton, M. J., & Donoghue, P. C. (2007). Paleontological evidence to date the tree of life. *Molecular Biology and Evolution*, *24*(1), 26–53.

Benton, M., & others. (2015). When life nearly died. Retrieved from http://www.publish.csiro.au/nid/276/pid/7621.htm

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., … Stadler, P. F. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, *69*(2), 313–319.

Boon, R. A., Jaé, N., Holdt, L., & Dimmeler, S. (2016). Long Noncoding RNAs: From Clinical Genetics to Therapeutic Targets? *Journal of the American College of Cardiology*, *67*(10), 1214–1226. https://doi.org/10.1016/j.jacc.2015.12.051

Botha, J., & Gaudin, T. (2007). An early pliocene pangolin (mammalia; pholidota) from langebaanweg, south africa. *Journal of Vertebrate Paleontology*, *27*(2), 484–491. https://doi.org/10.1671/0272-4634(2007)27[484:AEPPPF]2.0.CO;2

Bowden, R., MacFie, T. S., Myers, S., Hellenthal, G., Nerrienet, E., Bontrop, R. E., … Mundy, N. I. (2012). Genomic Tools for Evolution and Conservation in the

Chimpanzee: Pan troglodytes ellioti Is a Genetically Distinct Population. *PLoS Genetics*, *8*(3), e1002504. https://doi.org/10.1371/journal.pgen.1002504

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, İ., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *arXiv:1301.5406 [Q-Bio]*. Retrieved from http://arxiv.org/abs/1301.5406

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., … Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196. https://doi.org/10.1101/gr.6743907

Cantrell, M. A., Scott, L., Brown, C. J., Martinez, A. R., & Wichman, H. A. (2008). Loss of LINE-1 Activity in the Megabats. *Genetics*, *178*(1), 393–404. https://doi.org/10.1534/genetics.107.080275

Che, J., Wang, J., Su, W., Ye, J., Wang, Y., Nie, W., & Yang, F. (2007). Construction, characterization and FISH mapping of a bacterial artificial chromosome library of Chinese pangolin (Manis pentadactyla). *Cytogenetic and Genome Research*, *122*(1), 55–60.

Chen, M.-M., Li, Y., Chen, M., Wang, H., Li, Q., Xia, R.-X., … Qin, L. (2014). Complete mitochondrial genome of the atlas moth, Attacus atlas (Lepidoptera: Saturniidae) and the phylogenetic relationship of Saturniidae species. *Gene*, *545*(1), 95–101.

Chevreux, B. (2007). MIRA: an automated genome and EST assembler. Retrieved from http://archiv.ub.uni-heidelberg.de/volltextserver/7871/1/thesis_zusammenfassung.pdf

Cho, Y. S., Hu, L., Hou, H., Lee, H., Xu, J., Kwon, S., … Bhak, J. (2013). The tiger genome and comparative analysis with lion and snow leopard genomes. *Nature Communications*, *4*. https://doi.org/10.1038/ncomms3433

Choo, S. W., Rayko, M., Tan, T. K., Hari, R., Komissarov, A., Wee, W. Y., … Wong, G. J. (2016). Pangolin genomes and the evolution of mammalian scales and immunity. *Genome Research*, gr.203521.115. https://doi.org/10.1101/gr.203521.115

Clemson, C. M., Hutchinson, J. N., Sara, S. A., Ensminger, A. W., Fox, A. H., Chess, A., & Lawrence, J. B. (2009). An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Molecular Cell*, *33*(6), 717–726. https://doi.org/10.1016/j.molcel.2009.01.026

Coenye, T., Laevens, S., Willems, A., Ohlén, M., Hannant, W., Govan, J. R., … Vandamme, P. (2001). Burkholderia fungorum sp. nov. and Burkholderia caledonica sp. nov., two new species isolated from the environment, animals and human clinical samples. *International Journal of Systematic and Evolutionary Microbiology*, *51*(3), 1099–1107.

Collini, C. A. (1767). *Description Succincte du Cabinet d'Histoire Naturelle de Son Altesse Electorale Palatine ...*

Cuvier, G. (1833). *The animal kingdom: arranged in conformity with its organization*. G. & C. & H. Carvill. Retrieved from https://books.google.com.my/books?hl=en&lr=&id=fiJhAAAAIAAJ&oi=fnd&pg=PP6&dq=cuvier+The+animal+kingdom&ots=QMr5yo4xx5&sig=cwf0uJaN5pd1wr3dQX4_Bt0brA4

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R. F., … Hayes, B. J. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, *46*(8), 858–865. https://doi.org/10.1038/ng.3034

Davit-Béal, T., Tucker, A. S., & Sire, J.-Y. (2009). Loss of teeth and enamel in tetrapods: fossil record, genetic data and morphological adaptations. *Journal of Anatomy*, *214*(4), 477–501. https://doi.org/10.1111/j.1469-7580.2009.01060.x

Decker, J. E., Pires, J. C., Conant, G. C., McKay, S. D., Heaton, M. P., Chen, K., … Taylor, J. F. (2009). Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Sciences*, *106*(44), 18644–18649. https://doi.org/10.1073/pnas.0904691106

Delgado, S., Vidal, N., Veron, G., & Sire, J.-Y. (2008). Amelogenin, the major protein of tooth enamel: a new phylogenetic marker for ordinal mammal relationships. *Molecular Phylogenetics and Evolution*, *47*(2), 865–869.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., … Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*, *22*(9), 1775–1789. https://doi.org/10.1101/gr.132159.111

Du Toit, Z., Grobler, J. P., Kotzé, A., Jansen, R., Brettschneider, H., & Dalton, D. L. (2014). The complete mitochondrial genome of Temminck's ground pangolin (Smutsia temminckii; Smuts, 1832) and phylogenetic position of the Pholidota (Weber, 1904). *Gene*, *551*(1), 49–54.

Earl, D. A., Bradnam, K., John, J. S., Darling, A., Lin, D., Faas, J., … Paten, B. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, gr.126599.111. https://doi.org/10.1101/gr.126599.111

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

Ekblom, R., Farrell, L. L., Lank, D. B., & Burke, T. (2012). Gene expression divergence and nucleotide differentiation between males of different color morphs and mating

strategies in the ruff. *Ecology and Evolution*, *2*(10), 2485–2505. https://doi.org/10.1002/ece3.370

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042. https://doi.org/10.1111/eva.12178

Emerson, K. J., Merz, C. R., Catchen, J. M., Hohenlohe, P. A., Cresko, W. A., Bradshaw, W. E., & Holzapfel, C. M. (2010). Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences*, *107*(37), 16196–16200. https://doi.org/10.1073/pnas.1006538107

Emry, R. J., Skinner, M. F., & others. (1970). A North American Oligocene pangolin and other additions to the Pholidota. Bulletin of the AMNH; v. 142, article 6. Retrieved from http://digitallibrary.amnh.org/dspace/handle/2246/1078

Erickson, I. K., Cantrell, M. A., Scott, L., & Wichman, H. A. (2011). Retrofitting the Genome: L1 Extinction Follows Endogenous Retroviral Expansion in a Group of Muroid Rodents. *Journal of Virology*, *85*(23), 12315–12323. https://doi.org/10.1128/JVI.05180-11

Fatica, A., & Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nature Reviews Genetics*, *15*(1), 7–21.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … others. (2015). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, gkv1344.

Fischer, C., Koblmüller, S., Gülly, C., Schlötterer, C., Sturmbauer, C., & Thallinger, G. G. (2013). Complete Mitochondrial DNA Sequences of the Threadfin Cichlid (Petrochromis trewavasae) and the Blunthead Cichlid (Tropheus moorii) and Patterns of Mitochondrial Genome Evolution in Cichlid Fishes. *PLoS ONE*, *8*(6), e67048. https://doi.org/10.1371/journal.pone.0067048

Frésard, L., Leroux, S., Roux, P.-F., Klopp, C., Fabre, S., Esquerré, D., … Pitel, F. (2014). Genome-wide characterization of RNA editing in chicken: lack of evidence for non-A-to-I events. *bioRxiv*, 8912. https://doi.org/10.1101/008912

Gaiti, F., Fernandez-Valverde, S. L., Nakanishi, N., Calcino, A. D., Yanai, I., Tanurdzic, M., & Degnan, B. M. (2015). Dynamic and Widespread lncRNA Expression in a Sponge and the Origin of Animal Complexity. *Molecular Biology and Evolution*, *32*(9), 2367–2382. https://doi.org/10.1093/molbev/msv117

Gaubert, P., & Antunes, A. (2005). Assessing the taxonomic status of the Palawan pangolin Manis culionensis (Pholidota) using discrete morphological characters. *Journal of Mammalogy*, *86*(6), 1068–1074.

Gaudin, T. J. (2004). Phylogenetic relationships among sloths (Mammalia, Xenarthra, Tardigrada): the craniodental evidence. *Zoological Journal of the Linnean Society*, *140*(2), 255–305.

Gaudin, T. J., & Wible, J. R. (1999). The entotympanic of pangolins and the phylogeny of the Pholidota (Mammalia). *Journal of Mammalian Evolution*, *6*(1), 39–65.

Gavrieli, Y., Sherman, Y., & Ben-Sasson, S. A. (1992). Identification of programmed cell death in situ via specific labeling of nuclear DNA fragmentation. *The Journal of Cell Biology*, *119*(3), 493–501.

Ge, R.-L., Cai, Q., Shen, Y.-Y., San, A., Ma, L., Zhang, Y., … Wang, J. (2013). Draft genome sequence of the Tibetan antelope. *Nature Communications*, *4*, 1858. https://doi.org/10.1038/ncomms2860

Gebo, D. L., & Rasmussen, D. T. (1985). The Earliest Fossil Pangolin (Pholidota: Manidae) from Africa. *Journal of Mammalogy*, *66*(3), 538–541. https://doi.org/10.2307/1380929

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., … Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from

massively parallel sequence data. *Proceedings of the National Academy of Sciences*, *108*(4), 1513–1518. https://doi.org/10.1073/pnas.1017351108

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Gray, D. (1865). Revision of the Genera and Species of Entomophagous Edentata, Founded on the Examination of the Specimens in the British Museum. In *Proceedings of the Zoological Society of London* (Vol. 33, pp. 359–386). Wiley Online Library. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1469-7998.1865.tb02351.x/full

Gregory, T. R., Nicol, J. A., Tamm, H., Kullman, B., Kullman, K., Leitch, I. J., … Bennett, M. D. (2007). Eukaryotic genome size databases. *Nucleic Acids Research*, *35*(Database issue), D332–D338. https://doi.org/10.1093/nar/gkl828

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

HAGENBLAD, J., OLSSON, M., PARKER, H. G., OSTRANDER, E. A., & ELLEGREN, H. (2009). Population genomics of the inbred Scandinavian wolf. *Molecular Ecology*, *18*(7), 1341–1351. https://doi.org/10.1111/j.1365-294x.2009.04120.x

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research*, *41*(13), e129–e129.

HALBERT, N. D., WARD, T. J., SCHNABEL, R. D., TAYLOR, J. F., & DERR, J. N. (2005). Conservation genomics: disequilibrium mapping of domestic cattle

chromosomal segments in North American bison populations. *Molecular Ecology*, *14*(8), 2343–2362. https://doi.org/10.1111/j.1365-294x.2005.02591.x

Handbook of the Mammals of the World, Vol. 2: Hoofed Mammals. (2011). Barcelona: Lynx Edicions.

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, *22*(2), 160–174.

Hassanin, A., Hugot, J.-P., & van Vuuren, B. J. (2015). Comparison of mitochondrial genome sequences of pangolins (Mammalia, Pholidota). *Comptes Rendus Biologies*, *338*(4), 260–265.

Haussler, D., O'Brien, S. J., Ryder, O. A., Barker, F. K., Clamp, M., Crawford, A. J., … others. (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, *100*(6), 659–674.

He, Y., Zhang, Z., Peng, X., Wu, F., & Wang, J. (2013). De novo assembly methods for next generation sequencing data. *Tsinghua Science and Technology*, *18*(5), 500–514. https://doi.org/10.1109/TST.2013.6616523

Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., … Wilson, R. K. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, *432*(7018), 695–716. https://doi.org/10.1038/nature03154

Hofacker, I. L. (2009). RNA secondary structure analysis using the Vienna RNA package. *Current Protocols in Bioinformatics*, 12–2.

Hormozdiari, F., Konkel, M. K., Prado-Martinez, J., Chiatante, G., Herraez, I. H., Walker, J. A., … Eichler, E. E. (2013). Rates and patterns of great ape retrotransposition. *Proceedings of the National Academy of Sciences*, *110*(33), 13457–13462. https://doi.org/10.1073/pnas.1310914110

Hou, Y., & Lin, S. (2009). Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes. *PLoS ONE*, *4*(9). https://doi.org/10.1371/journal.pone.0006978

Hua, L., Gong, S., Wang, F., Li, W., Ge, Y., Li, X., & Hou, F. (2015). Captive breeding of pangolins: current status, problems and future prospects. *ZooKeys*, (507), 99–114. https://doi.org/10.3897/zookeys.507.6970

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, *26*(5), 680–682.

Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., … Flicek, P. (2009). Ensembl 2009. *Nucleic Acids Research*, *37*(Database issue), D690-697. https://doi.org/10.1093/nar/gkn828

Hughes, A. L., & Hughes, M. K. (1995). Small genomes for better flyers. *Nature*, *377*(6548), 391. https://doi.org/10.1038/377391a0

Hunter, S. S., Lyon, R. T., Sarver, B. A. J., Hardwick, K., Forney, L. J., & Settles, M. L. (2015). Assembly by Reduced Complexity (ARC): a hybrid approach for targeted assembly of homologous sequences. *bioRxiv*, 14662. https://doi.org/10.1101/014662

Jahr, S., Hentze, H., Englisch, S., Hardt, D., Fackelmayer, F. O., Hesch, R.-D., & Knippers, R. (2001). DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Research*, *61*(4), 1659–1665.

Jentink, F. A. (1903). Habits of the Scaly Anteater from Java. *Notes from the Leyden Museum*, *23*(4), 183–184.

John, J. A. S., Ekberg, J. A., Dando, S. J., Meedeniya, A. C., Horton, R. E., Batzloff, M., … others. (2014). Burkholderia pseudomallei penetrates the brain via destruction

of the olfactory and trigeminal nerves: implications for the pathogenesis of neurological melioidosis. *MBio*, *5*(2), e00025–14.

Johnson, R. (2012). Long non-coding RNAs in Huntington's disease neurodegeneration. *Neurobiology of Disease*, *46*(2), 245–254. https://doi.org/10.1016/j.nbd.2011.12.006

Johnson, W. E., Onorato, D. P., Roelke, M. E., Land, E. D., Cunningham, M., Belden, R. C., … O\textquotesingleBrien, S. J. (2010). Genetic Restoration of the Florida Panther. *Science*, *329*(5999), 1641–1645. https://doi.org/10.1126/science.1192891

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1–4), 462–467. https://doi.org/10.1159/000084979

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., … Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, gr.170720.113. https://doi.org/10.1101/gr.170720.113

Kannan, S., Chernikova, D., Rogozin, I. B., Poliakov, E., Managadze, D., Koonin, E. V., & Milanesi, L. (2015). Transposable element insertions in long intergenic non-coding RNA genes. *Frontiers in Bioengineering and Biotechnology*, *3*, 71.

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., … Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genetics*, *9*(4). https://doi.org/10.1371/journal.pgen.1003470

Katoh, K., Kuma, K., Toh, H., & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, *33*(2), 511–518.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S. L., & others. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, *14*(4), R36.

Kingdon, J. (1971). *East African Mammals: An Atlas of Evolution in Africa* (Vol. 1). Academic Press.

Knief, C. (2014). Analysis of plant microbe interactions in the era of next generation sequencing technologies. *Plant Genetics and Genomics*, *5*, 216. https://doi.org/10.3389/fpls.2014.00216

Kohn, M. H., Murphy, W. J., Ostrander, E. A., & Wayne, R. K. (2006). Genomics and conservation genetics. *Trends in Ecology & Evolution*, *21*(11), 629–637.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, *35*(suppl 2), W345–W349. https://doi.org/10.1093/nar/gkm391

Koonin, E. V., & Galperin, M. Y. (2003). *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Boston: Kluwer Academic. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK20260/

Kramerov, D. A., & Vassetzky, N. S. (2011). Origin and evolution of SINEs in eukaryotic genomes. *Heredity*, *107*, 487–495.

Krause, W. J., & Leeson, C. R. (1974). The stomach of the pangolin (Manis pentadactyla) with emphasis on the pyloric teeth. *Acta Anatomica*, *88*(1), 1–10.

Ksepka, D. T., Parham, J. F., Allman, J. F., Benton, M. J., Carrano, M. T., Cranston, K. A., … Warnock, R. C. M. (2015). The Fossil Calibration Database—A New Resource for Divergence Dating. *Systematic Biology*, *64*(5), 853–859. https://doi.org/10.1093/sysbio/syv025

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Li, A., Zhang, J., & Zhou, Z. (2014). PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics*, *15*, 311. https://doi.org/10.1186/1471-2105-15-311

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., … Wang, J. (2010). The sequence and de novo assembly of the giant panda genome. *Nature*, *463*(7279), 311–317. https://doi.org/10.1038/nature08696

Linnen, C. R., Kingsley, E. P., Jensen, J. D., & Hoekstra, H. E. (2009). On the Origin and Spread of an Adaptive Allele in Deer Mice. *Science*, *325*(5944), 1095–1098. https://doi.org/10.1126/science.1175826

Liu, H.-Q., Li, Y., Irwin, D. M., Zhang, Y.-P., & Wu, D.-D. (2014). Integrative analysis of young genes, positively selected genes and lncRNAs in the development of Drosophila melanogaster. *BMC Evolutionary Biology*, *14*(1), 241.

Liu, Y., Schröder, J., & Schmidt, B. (2013). Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics*, *29*(3), 308–315. https://doi.org/10.1093/bioinformatics/bts690

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., & Konstantinidis, K. T. (2012). Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLOS ONE*, *7*(2), e30087. https://doi.org/10.1371/journal.pone.0030087

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. https://doi.org/10.1186/2047-217X-1-18

Mäkinen, H., Vasemägi, A., McGinnity, P., Cross, T. F., & Primmer, C. R. (2015). Population genomic analyses of early-phase Atlantic Salmon (Salmo salar) domestication/captive breeding. *Evolutionary Applications*, *8*(1), 93–107. https://doi.org/10.1111/eva.12230

Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y., & Bryant, S. H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Research*, *30*(1), 281–283.

Margulies, E. H., Maduro, V. V. B., Thomas, P. J., Tomkins, J. P., Amemiya, C. T., Luo, M., & Green, E. D. (2005). Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(9), 3354–3359. https://doi.org/10.1073/pnas.0408539102

McCormack, J. E., Maley, J. M., Hird, S. M., Derryberry, E. P., Graves, G. R., & Brumfield, R. T. (2013). Corrigendum to "Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences" [Mol. Phylogenet. Evol. 62 (2012) 397–406]. *Molecular Phylogenetics and Evolution*, *66*(1), 440. https://doi.org/10.1016/j.ympev.2012.09.009

McKenna, M. C., Bell, S. K., & Simpson, G. G. (1997). *Classification of mammals above the species level*. Columbia University Press.

McKnight, D. A., & Fisher, L. W. (2009). Molecular evolution of dentin phosphoprotein among toothed and toothless animals. *BMC Evolutionary Biology*, *9*, 299. https://doi.org/10.1186/1471-2148-9-299

Merriman, B., R&D Team, I. T., & Rothberg, J. M. (2012). Progress in Ion Torrent semiconductor chip based sequencing. *ELECTROPHORESIS*, *33*(23), 3397–3417. https://doi.org/10.1002/elps.201200424

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly Algorithms for Next-Generation Sequencing Data. *Genomics*, *95*(6), 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, *12*, R112. https://doi.org/10.1186/gb-2011-12-11-r112

Moghadam, H. K., Pointer, M. A., Wright, A. E., Berlin, S., & Mank, J. E. (2012). W chromosome expression responds to female-specific selection. *Proceedings of the National Academy of Sciences*, *109*(21), 8207–8211. https://doi.org/10.1073/pnas.1202721109

Mohapatra, R. K., & Panda, S. (2014). Behavioural Descriptions of Indian Pangolins (*Manis crassicaudata*) in Captivity. *International Journal of Zoology*, *2014*, e795062. https://doi.org/10.1155/2014/795062

Mohapatra, R. K., Panda, S., Nair, M. V., & Acharjyo, L. N. (2015). Check list of parasites and bacteria recorded from pangolins (Manis sp.). *Journal of Parasitic Diseases*, 1–7. https://doi.org/10.1007/s12639-015-0653-5

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., & O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature*, *409*(6820), 614–618. https://doi.org/10.1038/35054550

Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*, *21*(suppl 2), ii79-ii85. https://doi.org/10.1093/bioinformatics/bti1114

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., … Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, *39*(13), e90–e90. https://doi.org/10.1093/nar/gkr344

Narzisi, G., & Mishra, B. (2011). Scoring-and-unfolding trimmed tree assembler: concepts, constructs and comparisons. *Bioinformatics (Oxford, England)*, *27*(2), 153–160. https://doi.org/10.1093/bioinformatics/btq646

Nguyen, P., Ma, J., Pei, D., Obert, C., Cheng, C., & Geiger, T. L. (2011). Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics*, *12*, 106. https://doi.org/10.1186/1471-2164-12-106

Nguyen, V. T., Leanne, C., & Tran, Q. P. (2010). Management Guidelines for Sunda pangolin (Manis javanica).

Nie, W., Wang, J., Su, W., Wang, Y., & Yang, F. (2009). Chromosomal rearrangements underlying karyotype differences between Chinese pangolin (<i>Manis pentadactyla</i>) and Malayan pangolin (<i>Manis javanica</i>) revealed by chromosome painting. *Chromosome Research*, *17*(3), 321–329. https://doi.org/10.1007/s10577-009-9027-0

Nisa, C. (2005). Morphological studies of the stomach of Malayan pangolin, Manis javanica. Retrieved from http://repository.ipb.ac.id/handle/123456789/299

NUNES, V. L., BEAUMONT, M. A., BUTLIN, R. K., & PAULO, O. S. (2010). Multiple approaches to detect outliers in a genome scan for selection in ocellated lizards (Lacerta lepida) along an environmental gradient. *Molecular Ecology*, *20*(2), 193–205. https://doi.org/10.1111/j.1365-294x.2010.04936.x

Ofusori, D. A., Caxton-Martins, E. A., Adenowo, T. K., Ojo, G. B., Falana, B. A., Komolafe, A. O., … Oluyemi, K. A. (2007). Morphometric study of the stomach of African pangolin (Manis tricuspis). *Sci. Res. Essays*, *2*(10), 465–61.

Otto, C., Hoffmann, S., Gorodkin, J., & Stadler, P. F. (2011). Fast local fragment chaining using sum-of-pair gap costs. *Algorithms for Molecular Biology : AMB*, *6*, 4. https://doi.org/10.1186/1748-7188-6-4

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics (Oxford, England)*, *23*(9), 1061–1067. https://doi.org/10.1093/bioinformatics/btm071

Pemberton, C. (2011). *English: Distribution of the Pangolins derived from constituent species maps.* Retrieved from https://commons.wikimedia.org/wiki/File:Manis_ranges.png

Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A. M., … Pecon-Slattery, J. (2011). A Molecular Phylogeny of Living Primates. *PLoS Genetics*, *7*(3), e1001342. https://doi.org/10.1371/journal.pgen.1001342

Pertoldi, C., Wójcik, J. M., Tokarska, M., Kawa\lko, A., Kristensen, T. N., Loeschcke, V., … Bendixen, C. (2009). Genome variability in European and American bison detected using the BovineSNP50 BeadChip. *Conservation Genetics*, *11*(2), 627–634. https://doi.org/10.1007/s10592-009-9977-y

Phillips, M. J. (2015). Four mammal fossil calibrations: balancing competing palaeontological and molecular considerations. *Palaeontologia Electronica*, *18*(1), 1–16.

Platt II, R. N., & Ray, D. A. (2012). A non-LTR retroelement extinction in Spermophilus tridecemlineatus. *Gene*, *500*(1), 47–53.

Pocock, R. I. (1924). The External Characters of: the Pangolins (Manid\a e). In *Proceedings of the Zoological Society of London* (Vol. 94, pp. 707–723). Wiley Online Library. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1096-3642.1924.tb03310.x/abstract

Prapong, T., Liumsiricharoen, M., Chungsamarnyart, N., Chantakru, S., Yatbantoong, N., Sujit, K., … others. (2009). Macroscopic and Microscopic Anatomy of Pangolinûs Tongue (Manis javanica). *Kasetsart Veterinarians*, 9.

Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J. R., Walenz, B., … Pääbo, S. (2012). The bonobo genome compared with the chimpanzee and human genomes. *Nature*, *486*(7404), 527–531. https://doi.org/10.1038/nature11128

Pryszcz, L. P., & Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, *44*(12), e113–e113. https://doi.org/10.1093/nar/gkw294

Qin, X.-M., Dou, S.-R., Guan, Q.-X., Qin, P.-S., & She, Y. (2012). Complete mitochondrial genome of the Manis pentadactyla (Pholidota, Manidae): Comparison of M. pentadactyla and M. tetradactyla. *Mitochondrial DNA*, *23*(1), 37–38. https://doi.org/10.3109/19401736.2011.643881

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Quinn, J. J., & Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, *17*(1), 47–62.

Rachmawati, A. F. (2011). Morfologi Organ Reproduksi Betina Trenggiling Jawa (Manis javanica) dengan Tinjauan Khusus pada Karakteristik Perkembangan Folikel dan Distribusi Karbohidrat pada Ovarium. Retrieved from http://repository.ipb.ac.id/handle/123456789/52318

R Core Team. (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013*. ISBN 3-900051-07-0. Retrieved from

http://scholar.google.com/scholar?cluster=8103611549594844363&hl=en&oi=scholarr

Rambaut, A. (2014). FigTree. Retrieved January 10, 2016, from http://tree.bio.ed.ac.uk/software/figtree/

Rambaut, A., Suchard, M., Xie, D., & Drummond, A. (2014). Tracer v1.6. Retrieved January 10, 2016, from http://beast.bio.ed.ac.uk/tracer

Roberts, A., Pimentel, H., Trapnell, C., & Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, *27*(17), 2325–2329. https://doi.org/10.1093/bioinformatics/btr355

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, *12*, R22. https://doi.org/10.1186/gb-2011-12-3-r22

Ronaghi, M., Uhlén, M., & Nyrén, P. (1998). A Sequencing Method Based on Real-Time Pyrophosphate. *Science*, *281*(5375), 363–365. https://doi.org/10.1126/science.281.5375.363

Ruhyana, A. Y. (2007). Kajian Morfologi Saluran Pernafasan Trenggiling (Manis javanica) dengan Tinjauan Khusus pada Trakea dan Paru-paru. Retrieved from http://repository.ipb.ac.id/handle/123456789/1730

Russello, M. A., Waterhouse, M. D., Etter, P. D., & Johnson, E. A. (2015). From promise to practice: pairing non-invasive sampling with genomics in conservation. *PeerJ*, *3*, e1106. https://doi.org/10.7717/peerj.1106

Salles, J. F., van Veen, J. A., & van Elsas, J. D. (2004). Multivariate Analyses of Burkholderia Species in Soil: Effect of Crop and Land Use History. *Applied and Environmental Microbiology*, *70*(7), 4012–4020. https://doi.org/10.1128/AEM.70.7.4012-4020.2004

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., … Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, *22*(3), 557–567. https://doi.org/10.1101/gr.131383.111

Sasaki, Y. T. F., Ideue, T., Sano, M., Mituyama, T., & Hirose, T. (2009). MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(8), 2525–2530. https://doi.org/10.1073/pnas.0807899106

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., … Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, *483*(7388), 169–175. https://doi.org/10.1038/nature10842

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864.

Semiadi, G., & others. (2013). Identifikasi Trenggiling (Manis javanica) Menggunakan Penanda Cytochrome B Mitokondria DNA (IDENTIFICATION OF PANGOLIN (MANIS JAVANICA DESMAREST, 1822) USING CYTOCHROME B mtDNA MARKER). *Jurnal Veteriner*, *14*(4). Retrieved from http://ojs.unud.ac.id/index.php/jvet/article/view/7682

Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H., & Wright, F. (1988). Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. *Nucleic Acids Research*, *16*(17), 8207–8211.

Sharp, P. M., Tuohy, T. M., & Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, *14*(13), 5125–5143.

Shen, L. (1996). *Shen Nong's herbal classic color atlas of Chinese Traditional medicine [M]*. *Beijing*. China Press of Traditional Chinese Medicine.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486

Simpson, J. T. (2013). Exploring Genome Characteristics and Sequence Quality Without a Reference. *arXiv:1307.8026 [Q-Bio]*. Retrieved from http://arxiv.org/abs/1307.8026

Simpson, J. T., & Durbin, R. (2010). Efficient construction of an assembly string graph using the FM-index. *Bioinformatics*, *26*(12), i367–i373. https://doi.org/10.1093/bioinformatics/btq217

Skinner, J. D., & Smithers, R. H. N. (1990). The mammals of the south African subregion. *University of Pretoria, Pretoria*.

Smit, A. F. A., Hubley, R., & Green, P. (1996). http:// www. repeatmasker. org. *RepeatMasker Open*, *3*, 1996–2004.

Stapley, J., Reger, J., Feulner, P. G., Smadja, C., Galindo, J., Ekblom, R., … Slate, J. (2010). Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, *25*(12), 705–712.

Storch, G. (1978). Eomanis waldi, ein Schuppentier aus dem Mittel-Eozän der "Grube Messel" bei Darmstadt (Mammalia: Pholidota). *Senckenbergiana Lethaea*, *59*(4–6), 503–529.

Storch, G., & Martin, T. (1994). Eomanis krebsi, ein neues Schuppentier aus dem Mittel-Eozän der Grube Messel bei Darmstadt (Mammalia: Pholidota). *Berliner Geowissenschaftliche Abhandlungen E*, *13*, 83–97.

Strandberg, J. (1918). A contribution to the question on the malformations of the ectoderme due to arrested deyelopment. *Nordiskt Medicinskt Arkiv*, *51*(1), 1–12.

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., … Zhao, Y. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research*, *41*(17), e166. https://doi.org/10.1093/nar/gkt646

Supek, F., & Vlahoviček, K. (2004). INCA: synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics*, *20*(14), 2329–2330.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725–2729.

Tarabeux, J., Zeitouni, B., Moncoutier, V., Tenreiro, H., Abidallah, K., Lair, S., … Houdayer, C. (2014). Streamlined ion torrent PGM-based diagnostics: BRCA1 and BRCA2 genes as a model. *European Journal of Human Genetics*, *22*(4), 535–541. https://doi.org/10.1038/ejhg.2013.181

Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., … Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, *22*(9), 1616–1625.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31*(1), 46–53. https://doi.org/10.1038/nbt.2450

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., … Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals

unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Ventura, M., Catacchio, C. R., Alkan, C., Marques-Bonet, T., Sajjadian, S., Graves, T. A., … Eichler, E. E. (2011). Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Research*, *21*(10), 1640–1649. https://doi.org/10.1101/gr.124461.111

Vezzi, F., Narzisi, G., & Mishra, B. (2012). Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. *PLoS ONE*, *7*(12), e52210. https://doi.org/10.1371/journal.pone.0052210

Voelkerding, K. V., Dames, S. A., & Durtschi, J. D. (2009). Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, *55*(4), 641–658. https://doi.org/10.1373/clinchem.2008.112789

vonHoldt, B. M., Pollinger, J. P., Earl, D. A., Knowles, J. C., Boyko, A. R., Parker, H., … Wayne, R. K. (2011). A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Research*, *21*(8), 1294–1305. https://doi.org/10.1101/gr.116301.110

WETZEL, D. P., STEWART, I. R. K., & WESTNEAT, D. F. (2011). Heterozygosity predicts clutch and egg size but not plasticity in a house sparrow population with no evidence of inbreeding. *Molecular Ecology*, *21*(2), 406–420. https://doi.org/10.1111/j.1365-294x.2011.05380.x

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., … others. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), 973–982.

WOLF, J. B. W., BAYER, T., HAUBOLD, B., SCHILHABEL, M., ROSENSTIEL, P., & TAUTZ, D. (2010). Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow

and its hybrid zone with the hooded crow. *Molecular Ecology*, *19*, 162–175. https://doi.org/10.1111/j.1365-294x.2009.04471.x

Wolstenholme, D. R. (1992). Animal mitochondrial DNA: structure and evolution. *International Review of Cytology*, *141*, 173–216.

Wong, P. B., Wiley, E. O., Johnson, W. E., Ryder, O. A., O'Brien, S. J., Haussler, D., … $author.lastName, $author firstName. (2012). Tissue sampling methods and standards for vertebrate genomics. *GigaScience*, *1*(1), 8. https://doi.org/10.1186/2047-217X-1-8

Wu, S., Liu, N., Zhang, Y., & Ma, G. Z. (2004). Assessment of threatened status of Chinese Pangolin (Manis pentadactyla). *Chinese Journal of Applied and Environmental Biology*, *10*(4), 456–461.

Wu, S., Liu, N., Zhang, Y., Ou, Z., & Chen, H. (2003). Measurement and comparison of skull variables in Chinese pangolin and Malayan pangolin. *Acta Theriologica Sinica*, *24*(3), 211–214.

Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*(9), 1859–1875.

Xu, R. S., Zong, X. H., & Li, X. G. (2009). [Controlled clinical trials of therapeutic effects of Chinese herbs promoting blood circulation and removing blood stasis on the treatment of reflex sympathetic dystrophy with type of stagnation of vital energy and blood stasis]. *Zhongguo Gu Shang= China Journal of Orthopaedics and Traumatology*, *22*(12), 920–922.

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329–342. https://doi.org/10.1038/nrg3174

Yang, C. W., Chen, S., Chang, C.-Y., Lin, M. F., Block, E., Lorentsen, R., … Dierenfeld, E. S. (2007). History and dietary husbandry of pangolins in captivity. *Zoo Biology*, *26*(3), 223–230.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., … Wang, J. (2010). Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science*, *329*(5987), 75–78. https://doi.org/10.1126/science.1190371

Yu, H.-T., Ma, G.-C., Lee, D.-J., Chin, S.-C., Chen, T.-L., Tsao, H.-S., … Chen, M. (2012). Use of a cytogenetic whole-genome comparison to resolve phylogenetic relationships among three species: Implications for mammalian systematics and conservation biology. *Theriogenology*, *77*(8), 1615–1623.

Yu, H.-T., Ma, G.-C., Lee, D.-J., Chin, S.-C., Tsao, H.-S., Wu, S.-H., … Chen, M. (2011). Molecular delineation of the Y-borne Sry gene in the Formosan pangolin (Manis pentadactyla pentadactyla) and its phylogenetic implications for Pholidota in extant mammals. *Theriogenology*, *75*(1), 55–64. https://doi.org/10.1016/j.theriogenology.2010.07.010

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

Zhang, L., Hua, N., & Sun, S. (2008). Wildlife trade, consumption and conservation awareness in southwest China. *Biodiversity and Conservation*, *17*(6), 1493–1516.

Zhang, Q., Chen, C.-Y., Yedavalli, V. S. R. K., & Jeang, K.-T. (2013). NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. *mBio*, *4*(1), e00596-512. https://doi.org/10.1128/mBio.00596-12

Zhang, Q., & Edwards, S. V. (2012). The Evolution of Intron Size in Amniotes: A Role for Powered Flight? *Genome Biology and Evolution*, *4*(10), 1033–1043. https://doi.org/10.1093/gbe/evs070

POSTERS & PUBLICATION

Posters

Ranjeev Hari, Aini Yasmin, Tan Joon Liang, Wee Wei Yee, Tan Tze King, Wong Guat Jah, and Choo Siew Woh."Assembly and Preliminary Characterisation of the first *Manis javanica* Genome". MSMBB Conference, Monash University, MALAYSIA. 2014

Mikhail Rayko, Aleksey Komossirov, Joon Liang Tan, Ranjeev Hari, Tan Tze King, Choo Siew Woh, Stephen J. O'Brien, Andrey Yurchenko. Genome Sequence and Annotation of Chinese Pangolin (*Manis pentadactyla*) and Malayan Pangolin (*Manis javanica*). Plant and Animal Genome Asia, SINGAPORE. 2015

Journal Publication

Hari R, Paterson IC, Siew Woh C. 2016. A new complete mitogenome of the critically endangered Chinese pangolin *Manis pentadactyla*. *Conservation Genetics Resources* 1–4.

Siew Woh C, Rayko M, Tze King T, **Hari R**, Komissarov A, Wei Yee W, Yurchenko A, Sergey K. *et al.* 2016. Pangolin genomes and the evolution of mammalian scales and immunity. *Genome Research*.

Aini MY, Tze King T, **Hari, R**, Klaus-Peter K, Wei Yee W, Agostinho A, Frankie TS, Jeffrine R-RJ, Kayalvizi K, Guat Jah W, et al. 2016. *De novo* sequencing, assembly and analysis of eight different transcriptomes from the Malayan pangolin. *Scientific Reports*